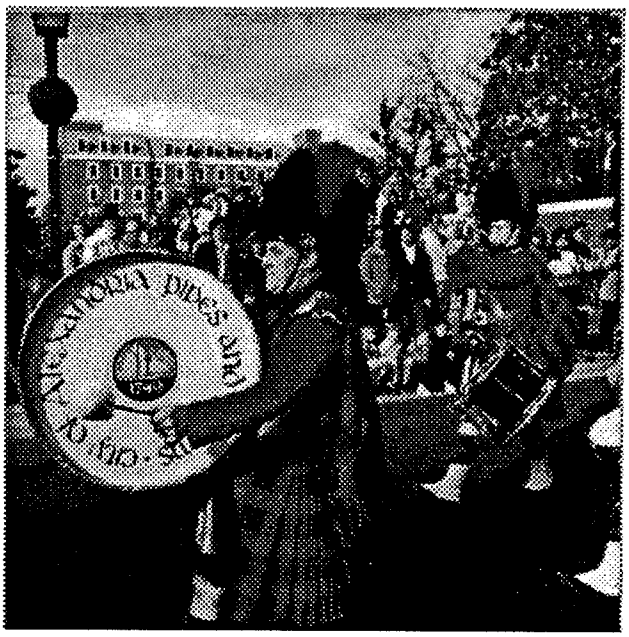
Proceedings

1994 IEEE-IMS Workshop on Information Theory and Statistics

October 27–29, 1994

Holiday Inn Old Town
Alexandria, Virginia, USA



Sponsored by
IEEE Information Theory Society
Institute of Mathematical Statistics

Proceedings

1994 IEEE-IMS Workshop on Information Theory and Statistics

October 27–29, 1994

Holiday Inn Old Town
Alexandria, Virginia, USA

Sponsored by
IEEE Information Theory Society
Institute of Mathematical Statistics

Proceedings 1994 IEEE-IMS Workshop on Information Theory and Statistics

Additional copies of this publication are available from:

IEEE Service Center
445 Hoes Lane
P.O.Box 1331
Piscataway, NJ 08855-1331, USA
1-800-678-IEEE

Cover photograph courtesy of Alexandria Convention & Visitors Bureau.

1994 IEEE-IMS
Workshop on Information Theory and Statistics

Foreword

The idea of a workshop on information theory and statistics was born from a recent series of exciting interactions between researchers from both communities covering a wide range of topics including data compression, complexity, Markov fields, large deviations, nonparametrics and sampling and wavelets. Our aim has been to provide a forum for the presentation and discussion of new technical results at the interface of these two disciplines. We hope that this Workshop will be the first of a series of many which will serve to promote a better understanding of the interplay between information theory and statistics.

Stamatis Cambanis
General Chairman

Prakash Narayan
Chairman, Technical Committee

Geoffrey C. Orsak
Chairman, Organizing Committee
Local Arrangements

Tom Fuja
Finance

Bernd-Peter Paris
Registration

Adrian Papamarcou
Publications

Acknowledgments

We wish to thank the Office of Naval Research, the National Security Agency and the National Science Foundation for their generous financial support of the Workshop. We also wish to express our gratitude to Dr. Julia Abrahams for her sustained encouragement and invaluable advice; to Ms. Denise Best, whose enthusiasm, patience and skill were instrumental to the realization of this Workshop; and to our graduate students Angelos Kanlis, Sanjeev Khudanpur, Nol Rananand and Ramin Rezaiifar who volunteered to toil for this worthy cause.

TECHNICAL PROGRAM

Wednesday, October 26

7:00-9:00 pm: WELCOME RECEPTION

Thursday, October 27

PLENARY TALKS, INVITED SESSIONS AND SPECIAL TALK

8:00–9:00 am: PLENARY TALK

T. M. Cover, *Stanford U.,* "Information theory and statistics."

9:00 am–12:15 pm: SESSION I

Stochastic Complexity and Universal Data Compression

Session Organizers: P. Shields, U. of Toledo, and J. Ziv, Technion

9:00-9:30	**I. Csiszár,** *Hungarian Academy of Sciences,* "Maximum entropy and related methods."
9:30-10:00	**N. Merhav,** *Technion,* "A stronger version of the redundancy-capacity theorem of universal coding," (joint work with **M. Feder**).
10:00-10:30	**J. Ziv,** *Technion,* "Bounds on universal coding: The next generation," (joint work with **Y. Hershkovits.**)
10:30-10:45	Coffee Break
10:45-11:15	**B. Clarke,** *U. of British Columbia,* "Jeffreys' prior yields the asymptotic minimax redundancy," (joint work with **A.R. Barron**).
11:15-11:45	**B. Yu,** *U. of California at Berkeley,* "Lower bounds on expected redundancy."
11:45-12:15	**P. Shields,** *U. of Toledo,* "When is the weak rate equal to the strong rate?"

1:45 –2:45 pm: PLENARY TALK

R. M. Gray, *Stanford U.,* "Bayes risk-weighted vector quantization."

2:45 pm–6:00 pm: SESSION II

Vector Quantization, Classification and Regression Trees

Session Organizers: R. M. Gray, Stanford U., and R. Olshen, Stanford U.

2:45-3:15	**R. Olshen,** *Stanford U.,* "Variable-rate, lossy, tree-structured codes and digital radiography."
3:15-3:45	**A. Nobel,** *U. of Illinois/U. of N. Carolina,* "Greedy growing of tree-structured classification rules using a composite splitting criterion."
3:45-4:15	**R. Picard,** *MIT,* "Tree-structured clustered probability models for texture," (joint work with **K. Popat**).
4:15-4:30	Coffee Break
4:30-5:00	**Q. Xie,** *U. of British Columbia,* "Nonparametric classifier design using vector quantization," (joint work with **R.K. Ward and C.A. Laszlo**).
5:00-5:30	**M. Riley,** *AT&T,* "Tree-based models for speech and language."
5:30-6:00	**X. Wu,** *U. of Western Ontario,* "Image coding via bintree segmentation and texture VQ."

7:30–8:15 pm: SPECIAL TALK

H.V. Poor, *Princeton U.,* "Maximum entropy and robust prediction on a simplex."

TECHNICAL PROGRAM (cont.)

Friday, October 28

PLENARY TALKS, INVITED SESSIONS AND SPECIAL TALK

8:00–9:00 am: PLENARY TALK

J. Rissanen, *IBM San Jose,* "Fisher information, stochastic complexity, and universal modeling."

9:00 am–12:15 pm: SESSION III

Randomization Complexity and Information Theory

Session Organizers: M. Burnashev, IPPI, Moscow, and S. Verdú, Princeton U.

9:00-9:30 **R.J. Lipton,** *Princeton U.,* "Coding for noisy feasible channels."

9:30-10:00 **L. Schulman,** *U. of California at Berkeley,* "Coding for distributed computation."

10:00-10:30 **S. Verdú,** *Princeton U.,* "Minimal randomness and information theory."

10:30-10:45 Coffee Break

10:45-11:15 **Y. Steinberg,** *George Mason U.,* "Finite precision intrinsic randomness and source resolvability," (joint work with **S. Verdú**).

11:15-11:45 **Z. Zhang,** *U. of Southern California,* "Identification via compressed data," (joint work with **R. Ahlswede and E.-H. Yang**).

11:45-12:15 **M. Burnashev,** *IPPI, Moscow,* "Testing of composite hypotheses and ID-codes." (joint work with **S. Verdú**).

1:45 –2:45 pm: PLENARY TALK

M. Vetterli, *U. of California at Berkeley,* "Signal expansions for compression."

2:45 pm–6:00 pm: SESSION IV

Nonparametric Function Estimation

Session Organizers: E. Masry, U. of California at San Diego, and I. Johnstone, Stanford U.

2:45-3:15 **A. Barron,** *Yale U.,* "Asymptotically optimal model selection and neural nets."

3:15-3:45 **R. Khasminskii,** *Wayne State U.,* "Some estimation problems in infinite dimensional Gaussian white noise," (joint work with **I. Ibragimov**).

3:45-4:15 **E. Masry,** *U. of California at San Diego* "Local polynomial estimation of regression functions for mixing processes," (joint work with **J. Fan**).

4:15-4:30 Coffee Break

4:30-5:00 **L. Breiman,** *U. of California at Berkeley,* "Issues and advances in high-dimensional function estimation."

5:00-5:30 **L. Györfi,** *Tech. Univ., Budapest,* "The asymptotic normality of global errors for a histogram-based density estimate."

5:30-6:00 **P. Hall,** *Australian National U.,* "Bandwidth choice and convergence rates in density estimation with long-range dependent data," (joint work with **S.N. Lahiri and Y.K. Truong**).

7:30–8:15 pm: SPECIAL TALK

A. Dembo, *Stanford U.,* "Large deviations in information theory and statistics."

TECHNICAL PROGRAM (cont.)

Saturday, October 29

PLENARY TALK AND INVITED SESSIONS

9:00 am–12:15 pm: SESSION V

Markov Random Fields

Session Organizers: M. Miller, Washington U., and L. Pitt, U. of Virginia

9:00-9:30	**F. Comets,** *U. of Paris,* "Large deviations and consistent estimates for Gibbs random fields."
9:30-10:00	**Y. Amit,** *U. of Chicago,* "Large deviations and the rate distortion theorem for Gibbs distributions."
10:00-10:30	**L.D. Pitt,** *U. of Virginia,* "Estimation and prediction for (mostly Gaussian) Markov fields in the continuum."
10:30-10:45	Coffee Break
10:45-11:15	**J.S. Rosenthal,** *U. of Toronto,* "Markov chain Monte Carlo algorithms."
11:15-11:45	**J.A. O'Sullivan,** *Washington U.,* "Markov random fields on graphs for natural languages," (joint work with **K. Mark and M.I. Miller**).
11:45-12:15	**B. Hajek,** *U. of Illinois,* "Equilibria in infinite random graphs."

1:45 –2:45 pm: PLENARY TALK

D. Geman, *U. of Massachusetts,* " The entropy strategy for shape recognition."

2:45 –6:00 pm: SESSION VI

Theory and Applications of Wavelets

Session Organizers: D. Donoho, Stanford U., and S. Mallat, NYU

2:45-3:15	**R. Coifman,** *Yale U.,* "Selection of best bases for classification and regression," (joint work with **N. Saito**).
3:15-3:45	**R.A. DeVore,** *U. of South Carolina,* "The role of approximation and smoothness spaces in compression and noise removal," (joint work with **V. Temlyakov**).
3:45-4:15	**D. Donoho,** *Stanford U.,* "Adaptive signal representations: How much adaptation is too much?"
4:15-4:30	Coffee Break
4:30-5:00	**P. Flandrin,** *ENS-Lyon,* "Tracking long-range dependencies with wavelets," (joint work with **P. Abry**).
5:00-5:30	**S. Mallat,** *NYU,* "Wavelet vector quantization with matching pursuit," (joint work with **G. Davis**).
5:30-6:00	**A.S. Willsky,** *MIT,* "Multiresolution models for random fields and their use in statistical image processing," (joint work with **H. Krim and W.C. Karl**).

TECHNICAL PROGRAM (cont.)

Thursday, October 27

8:30–10:30 pm: POSTER SESSION I, Reception

1. **G. Cheang,** *Yale U.*, "Neural network approximation and estimation of functions."
2. **J.W. Craig,** *Interstate Electronics Corp.,* "Markov chains and random walks in data communication receivers."
3. **P.M. Djurić,** *SUNY Stony Brook,* "MMSE parameter estimation of exponentially damped sinusoids," (joint work with **H.-T. Li**).
4. **M.A.T. Figueiredo,** *Instituto Técnico PORTUGAL,* "Adaptive edge detection in compound Gauss-Markov random fields using the minimum description length principle," (joint work with **J.M.N. Leitão**).
5. **R.L. Fry,** *Johns Hopkins U.,* "Maximized mutual information using macrocanonical probability distributions."
6. **S. Goswami,** *Carnegie Mellon U.,* "Sample path description of Gauss Markov random fields," (joint work with **J.M.F. Moura**).
7. **H.J. Holz,** *George Washington U.,* "Non-Parametric discriminatory power," (joint work with **M.H. Loew**).
8. **R.E. Krichevskii,** *Mathematical Institute and State U., Novosibirsk, Russia,* "Shannon-Hartley entropy ratio under Zipf law," (joint work with **M.P. Scharova**).
9. **A. Lapidoth,** *Stanford U.,* "Mismatched encoding in rate distortion theory"
10. **M.B. Maljutov,** *Moscow U.,* "On the mean rate of sequential search for significant variables of a function in noise."
11. **R. Matzner,** *Federal Armed Forces U. Munich,* "SNR estimation and blind equalization (deconvolution) using the kurtosis," (joint work with **K. Letsch**).
12. **D.S. Modha,** *U. Cal. at San Diego,* "Minimum complexity regression estimation with weakly dependent observations," (joint work with **E. Masry**).
13. **A.T. Murgan,** *U. of Politehnica, Bucharest,* "A comparison of algorithms for lossless data compression using the Lempel-Ziv-Welch type method," (joint work with **R. Radescu**).
14. **L.B. Nelson,** *Princeton U.,* "EM and SAGE algorithms for multi-user detection," (joint work with **H.V. Poor**)
15. **M. Pawlak,** *U. of Ulm, Germany,* "Nonparametric estimation of a class of smooth functions," (joint work with **U. Stadtmüller**)
16. **S.E. Posner,** *Princeton U.,* "Consistency and rates of convergence of k_n nearest neighbor estimation under arbitrary sampling," (joint work with **S.R. Kulkarni**).
17. **W.L. Poston,** *Naval Surface Warfare Ctr.,* "Choosing data sets that optimize the determinant of the Fisher information matrix," (joint work with **J.L. Solka**).
18. **J. Solka,** *Naval Surface Warfare Ctr.,* "The application of Akaike information criterion based pruning to nonparametric density estimates," (joint work with **C. Priebe, G. Rogers, W. Poston and D. Marchette**).
19. **Y. Steinberg,** *George Mason U.,* "Improved Ziv-Zakai lower bound for vector parameter estimation," (joint work with **K.L. Bell, Y. Ephraim, and H.L. Van Trees**).
20. **B.G. Talbot,** "Source coding with a reversible memory-binding probability density transformation," (joint work with **L.M. Talbot**).
21. **Z. Tian,** *Northwestern Polytechnical U., China,* "Projection pursuit autoregression and projection pursuit moving average."

TECHNICAL PROGRAM (cont.)

POSTER SESSION I (cont.)

22. **E.C. van der Meulen,** *Katholieke U. Leuven, Belgium,* "Root-n consistent estimators of entropy for densities with unbounded support," (joint work with **A.B. Tsybakov**).
23. **N. Warke,** *George Mason U.,* "On the theory and application of universal classification to signal detection," (joint work with **G.C. Orsak**).
24. **R. Zamir,** *Cornell U.,* "A matrix form of the Brunn-Minkowski inequality and geometric rates," (joint work with **M. Feder**).

Friday, October 28

8:30–10:30 pm: POSTER SESSION II, Reception

1. **J.O. Chapa,** *Rochester Inst. of Tech.,* "Constructing wavelets from desired signal functions," (joint work with **M. Raghuveer**).
2. **J. Chen,** *U. Electro-Comm., Japan,* "Wavelet transform based ECG data compression with desired reconstruction signal quality." (joint work with **S. Itoh** and **T. Hashimoto**).
3. **W. Duanyi,** *Beijing U., China* "Application of Markov model in mobile communication channel." (joint work with **H. Zhengming**).
4. **R. Eier,** *Tech. U. of Vienna, Austria* "Markov chains for modeling and analyzing digital data signals."
5. **M. Foodeei,** *U. du Quebec, Canada* "Quantization theory and EC-CELP advantages at low bit rates." (joint work with **E. Dubois**).
6. **R.L. Fry,** *Johns Hopkins U.,* "Neural processing of information."
7. **J. Kogan,** *NYU,* "The most informative stopping times for Viterbi algorithm: sequential properties."
8. **S. Krishnamachari,** *U. of Maryland,* "Modeling Gauss Markov random fields at multiple resolutions," (joint work with **R. Chellappa**).
9. **D.E. Lake,** *ONR,* "Detecting regularity in point processes generated by humans."
10. **T.H. Li,** *Texas A&M U.,* "New distortion measures for speech processing," (joint work with **J.D. Gibson**).
11. **Z. Li,** *Polytechnic U., China,* "Nonparametric kernel estimation for the error density." (joint work with **S.Z. Zou**).
12. **O. Mayora-Ibarra,** *Inst. Tech. de Estudios Sup. de Monterrey,* "Neural networks for error correction of Hamming codes," (joint work with **A. González-Gutiérrez, J.C. Ruiz-Suárez**).
13. **I.S. Moskowitz,** *Naval Research Lab.,* "Discussion of a statistical channel," (joint work with **M.H. Kang**).
14. **F. Müller,** *Aachen U. of Tech., Germany,* "Asymptotic performance evaluation of mismatched vector quantizers using sub-Gaussian sources."
15. **T. Robert,** *ENSEEIHT/GAPSE, France,* "Continuously evolving classification of signals corrupted by an abrupt change," (joint work with **J.Y. Tourneret**).
16. **R.R. Snapp,** *U. of Vermont,* "The finite-sample risk of the k-nearest-neighbor classifier under the L_p metric," (joint work with **S.S. Venkatesh**).
17. **L.M. Talbot,** "Characteristics of a statistical fuzzy grade-of-membership model in the context of unsupervised data clustering, (joint work with **H.D. Tolley, B.G. Talbot, H.D.Mecham**).

TECHNICAL PROGRAM (cont.)

POSTER SESSION II (cont.)

18. **Y. Wang,** *U. of Missouri at Columbia,* "Function estimation via wavelets for data with long-range dependence."
19. **Y. Wang,** *U. of Maryland Baltimore County,* "Unsupervised medical image analysis by multiscale FNM modeling and MRF relaxation labeling," (joint work with **T. Adali, and T. Lei**).
20. **M.A. Wickert,** *U. of Colorado,* "An additive congruential method for generating a multiple occurrence uniform random sequence," (joint with **D.M. Ionescu**).
21. **Y. Yang,** *Yale U.,* "An asymptotic property of model selection criteria."
22. **J. Zhang,** *U. of Wisconsin-Milwaukee,* "Wavelet networks for functional learning," (joint work with **G.G. Walter**).

TABLE OF CONTENTS

In each category, papers appear in order of presentation (see Technical Program).

PLENARY & SPECIAL TALKS

Information Theory and Statistics

Tom Cover[1]

Departments of Electrical Engineering and Statistics, Stanford University, Durand 121, Stanford, CA 94305-4055, USA,
email: cover@isl.stanford.edu

Abstract — **The main theorems in information theory and statistics are put in context, the differences are discussed, and some of the open research problems are mentioned.**

I. Introduction

Probability theory has produced a number of strong general statements — truths about stochastics processes that give random processes a deterministic flavor. These successes include the strong law of large numbers, the central limit theorem, the law of the iterated logarithm, the ergodic theorem, and limit theorems for Markov processes.

Information theory, on the other hand, has been primarily motivated by an attempt to optimize certain processes, for example, to minimize the description length of random processes or to maximize the number of distinguishable signals in the presence of noise. This different orientation — optimization — has led to a number of additional insights which contribute to the body of knowledge in probability theory. For example, the central limit theorem can be proved by way of the entropy power inequality, yielding a monotonic convergence to the limit. And the law of large numbers has a counterpart in the asymptotic equipartition property, which says that all ergodic stochastic processes can be considered as a uniform distribution over a small set of typical sequences characterized by the entropy rate.

II. Specifics

We will demonstrate some of the points of intersection of information theory and statistics, and mention some problems in physics and computer science that require a rigorous probabilistic treatment.

The discussion will include the following:

1. Chernoff information, error exponents, large deviation theory.
2. The geometry of information.
3. Structure of ergodic processes, the AEP and the Slepian Wolf theorem.
4. The common basis for the Cramer-Rao, entropy power, Brunn-Minkowski, and Heisenberg uncertainty inequalities. (See Dembo.)
5. Entropy rate (compressibility limits), channel capacity (distinguishability limits). The duality of the two.
6. The central limit theorem and the entropy power inequality. (See Barron.)
7. Information loss and the second law. The argument that entropy will be lost when mass is thrown into a black hole, together with the even stronger belief that entropy increases (the second law of thermodynamics), led Beckenstein and Hawking to argue that the mass of the black hole (which increases when matter is thrown into it) is proportional to its entropy (the logarithm of the number of ways in which it could be made) thus preserving the second law.
8. Entropy increase. The $\dot{H}$ theorem in statistical mechanics shows that entropy increases with time. But the laws of physics are time reversible. What is going on?
9. Investment processes. Duality with data compression.

III. Remarks

Certain theorems from information theory like the asymptotic equipartition property (the Shannon-MacMillan-Breiman theorem) may deserve to be considered part of the hard core of probability theory. Yet other results in information theory like the entropy power inequality turn out to play an important role in interpreting the central limit theorem. And finally, some of the tools in information theory may have important roles to play in physics, just as ergodic theory, developed in the 1930s, resolved some of the problems in statistical mechanics.

Acknowledgements

The work of Imre Csiszar has greatly influenced my understanding.

References

[1] I. Csiszar, T. Cover and B.S. Choi. Conditional Limit Theorems under Markov Conditioning. *IEEE Transactions on Information Theory*, IT-33(6): 788-801, November 1987.

[2] A. Dembo, T.Cover and J. Thomas. Information Theoretic Inequalities. *IEEE Transactions on Information Theory*, 37(6):1501-1518, November 1991.

[3] A. Barron. Entropy and the Central Limit Theorem. *The Annals of Probability*, 14(1):336-342, 1986.

[1]This work was partially supported by NSF Grant NCR-9205663 and JSEP Contract DAAH04-94-G-0058.

Bayes risk-weighted vector quantization

Robert M. Gray[1]
Department of Electrical Engineering, Stanford, CA 94305-4055

Abstract — **Lossy compression and classification algorithms both attempt to reduce a large collection of possible observations into a few representative catogories so as to preserve essential information. In this talk a recently developed framework for combining classification and compression into one or two quantizers is described along with some examples and related to other quantizer-based classification schemes.**

The object of classification is to map an observed vector into one of a finite collection of indices representing a class or type of data. For example, one might view a block of pixels in a digital mammogram and wish to classify the block as containing a microcalcification or not. The quality of of the classifier is typically measured by its average Bayes risk. Another important attribute of the classifier is its complexity, how hard it is to convert the observed vector into the final decision. The object of vector quantization is to map an observed vector into one of a finite number of representatives or templates. Here quality is typically measured by an average distortion such as squared error. Instead of complexity, the second parameter of primary importance is typically bit rate, measured either by the log of the number of templates or by the entropy of the quantizer output.

Classification and compression can both be viewed as a quantization operation, mapping a possibly continuous space into a finite one. The measures of quality differ, but both Bayes risk and squared error can be viewed as a form of distortion to be minimized subject to constraints on complexity or bit rate. Futhermore, bit rate is relevant to classification if continuous data is quantized to prior to digital classification, and complexity is important for compression to ensure efficient software or hardware implementation.

Many quantizer-based classifiers have been proposed in the literature, including classical nearest neighbor and clustered variations [1, 2, 3]. Perhaps the most famous quantization approach to classification is Kohonen's "learning vector quantizer" (LVQ) [4]. While codebook design differs, all use a Euclidean nearest neighbor encoder.

An alternative approach is to incorporate explicitly a Bayes risk term into the average distortion minimized by a quantizer using a Lagrange multiplier and thereby include both a distortion term reflecting the general quality of the reproduction (such as signal-to-noise ratio) and one reflecting the intended application (such as Bayes risk or classification error). By weighting these two components one can, in effect, optimize for general appearance and specific task [5, 6, 7, 8, 9, 10].

Let q be a k-dimensional vector quantizer with codebook $\mathcal{C}$, partition $\mathcal{P}$, encoder α, and decoder β. Let δ be a classifier assigning a class label $\delta(i) \in \mathcal{H}$ to each possible encoder output $i = 1, \ldots, N$, producing an overall classification rule of $\gamma(x) = \delta(\alpha(x))$. Let d denote a distortion measure such as squared error.

The compression performance measured by mean squared error is

$$D(\alpha, \beta) = \sum_{i=1}^{N} E[d(X, \beta(i))|\alpha(X) = i] \Pr(\alpha(X) = i). \quad (1)$$

The classification performance measured by Bayes risk is

$$B(\alpha, \delta) = \sum_{k=1}^{M} \sum_{j=1}^{M} C_{jk} \Pr(\delta(\alpha(X)) = k \text{ and } Y = j) \quad (2)$$

In order to simultaneously consider the compression and classification abilities of the encoder, we use a Lagrangian modified distortion expression which includes both ordinary distortion and classification error:

$$J_\lambda(\alpha, \beta, \delta) = D(\alpha, \beta) + \lambda B(\alpha, \delta). \quad (3)$$

This formulation leads to necessary conditions for an optimal code and a generalized Lloyd iterative code design algorithm, which are surveyed with examples in this talk.

[1]Portions of this work was supported in part by the National Science Foundation under grants MIP-9311190, and DMS-9101548 and by the National Institutes of Health under grant 1RO1-CA55325.

References

[1] Q. Xie, C. A. Laszlo, and R. K. Ward, "Vector quantization technique for nonparametric classifier design", *IEEE Trans. Pattern Anal. and Mach. Int.*, vol. 15, no. 12, pp. 1326–1330, Dec. 1993.

[2] K. Popat and R. W. Picard, "Novel cluster-based probability model for texture synthesis, classification, and compression", in *Proc. SPIE Visual Communications and Image Processing*, Boston, MA, Nov. 1993.

[3] K. Popat and R. W. Picard, "Cluster-based probability model applied to image restoration and compression", in *PICASSP*, Adelaide, Australia, April 1994.

[4] T. Kohonen, *Self-organization and associative memory*, Springer-Verlag, Berlin, third edition, 1989.

[5] K. L. Oehler, P. C. Cosman, R. M. Gray, and J. May, "Classification using vector quantization", in *Conference record of the Twenty-Fifth Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1991, pp. 439–445.

[6] K.L. Oehler and R.M. Gray, "Combining image classification and image compression using vector quantization", in *Proceedings of the 1993 IEEE Data Compression Conference (DCC)*, J.A. Storer and M. Cohn, Eds., Snowbird, Utah, March 1993, pp. 2–11, IEEE Computer Society Press.

[7] K. L. Oehler and R. M. Gray, "Combining image compression and classification using vector quantization," *IEEE Transactions PAMI*, to appear.

[8] K. Perlmutter, R. M. Gray, K. L. Oehler, and R. A. Olshen, "Bayes risk weighted tree structured vector quantization with estimated class posteriors", in *Proceedings Data Compression Conference*, Snowbird, Utah, April 1993, pp. 274–283.

[9] Rick D. Wesel and R. M. Gray, "Bayes risk weighted VQ and learning VQ", in *Proceedings IEEE Data Compression Conference*, Snowbird, Utah, April 1994.

[10] K. O. Perlmutter, R. M. Gray, K. L. Oehler, and R. A. Olshen, "Bayes risk weighted vector quantization with estimated class posteriors", Submitted for possible publiation, 1994.

Maximum Entropy and Robust Prediction on a Simplex

H. Vincent Poor[1]

Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA

Abstract — **The related problems of (finite-length) robust prediction and maximizing spectral entropy over a simplex of covariance matrices are considered. General properties of iterative solutions of these problems are developed, and monotone convergence proofs are presented for two algorithms that provide such solutions. The analogous problems for simplexes of spectral densities are also considered.**

I. Summary

The problem of designing an optimum predictor for an observed time series is, of course, a fundamental one that arises in innumerable applications. One reason for the central role of this problem is that the design of such a predictor is tantamount to the selection of a stochastic realization model of the time series [6], which can in turn be used in tasks such as control, data compression, and so forth.

The classical Levinson problem - that is the finite-length linear prediction of a covariance-stationary time series - is a central problem within this general area. The maximum-entropy spectrum fitting problem [8] is the counterpart of the Levinson problem in the context of stochastic realization. Both the Levinson problem and the maximum-entropy spectrum fitting problem involve the computation of a model/predictor for a time series from knowledge of the covariance structure of the series up to some finite lag, say p.

When this covariance structure is not known exactly, but rather is known only to lie in an uncertainty class of covariances, then the classical Levinson formulation for predictor design can be replaced by a minimax robustness formulation, as developed in several works (see [9] for a review). In this minimax formulation, the robust predictor is the optimum predictor designed for a least-favorable covariance structure, chosen to maximize the spectral entropy in the time series. In the context of model determination, the counterpart to robust prediction is approximate stochastic realization [5].

In this talk, we consider the minimax robust prediction problem for the situation in which the uncertainty class of covariances is a finite-dimensional simplex of covariance matrices. As we shall note, this formulation arises naturally from the consideration of confidence intervals for covariance estimates. Moreover, solutions for such uncertainty classes can be used as intermediate iterations for other convex uncertainty classes, as will be discussed in the paper (see also [3]).

This talk is organized as follows (details of this work can be found in [10]). First, we review briefly the problems of finite-length minimax robust prediction and maximum-entropy spectrum fitting, and in particular we note that both problems have essentially the same solution. We also provide necessary and sufficient conditions for solutions to these problems over general uncertainty classes and over simplexes. Next, general properties of iterative solutions to these problems are presented. In particular, the convergence of a series of entropies to the maximum entropy is shown to be equivalent to convergence of the corresponding covariances and predictors. Two iterative algorithms for maximizing entropy over a simplex are then developed, together with proofs of their monotone convergence. One of these algorithms generalizes Nelson's algorithm for solving mimimax decision problems [7], and the other is similar in approach to the Arimoto-Blahut algorithm of information theory [1, 2]. Finally, we consider the analogous problem for infinite-length prediction (i.e., $p = \infty$), in which the covariance structure is specified in terms of an (uncertain) spectral density. Results analogous to those of for the finite-length case are developed, and it is noted that this infinite-length version of the problem is identical mathematically to an optimization problem arising in portfolio theory [4].

[1]This work was supported by a UK SERC Senior Visiting Fellowship at Imperial College, London.

Acknowledgements

The author is grateful to Mark H. A. Davis and Matthieu Biron of Imperial College, London, for stimulating discussions that led to the consideration of the problem treated in this paper.

References

[1] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inform. Theory* **IT-18** (1) 14 - 20, 1972.

[2] R. E. Blahut, "Computation of channel capacity and rate-distortion functions" *IEEE Trans. Inform. Theory* **IT-18** (4) 460 - 473, 1972.

[3] C.-I Chang and L. D. Davisson, "Two iterative algorithms for finding minimax solutions," *IEEE Trans. Inform. Theory* **36** (1) 126 - 140, 1990.

[4] T. Cover, "An algorithm for maximizing expected log investment return," *IEEE Trans. Inform. Theory* **IT-30** (2) 369 - 373, 1984.

[5] M. H. A. Davis and P. G. Fotopoulos, "On the approximate stochastic realisation problem," *Proc. 30th IEEE CDC*, Brighton, UK, December 1991, 1698 - 1699.

[6] M. H. A. Davis and R. B. Vinter, *Stochastic Modeling and Control.* (Chapman and Hall: London, 1986)

[7] W. Nelson, "Minimax solution of statistical decision problems by iteration," *Ann. Math. Stat.* **37** 1643 - 1657, 1966.

[8] A. Papoulis, "Maximum entropy and spectral estimation: A review," *IEEE Trans. Acoust., Speech, Sig. Proc.* **ASSP-29** (6) 1176 - 1186, 1981.

[9] H. V. Poor, "Nonstandard Methods in Prediction," in *Adaptive Signal Processing*, L. D. Davisson and G. Longo, Eds. (Springer-Verlag: Vienna, 1991) pp. 173-203.

[10] H. V. Poor, *Maximum Entropy and Robust Prediction on a Simplex.* Technical Report No. B94/24, Centre for Process Systems Engineering, Imperial College of Science, Technology & Medicine, London University, UK, 1994.

Fisher Information, Stochastic Complexity, and Universal Modeling

J. Rissanen

IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, Ca 95120-6099

Abstract — The main objective in universal modeling is to construct a process for a class of model processes which for long strings, generated by any of the models in the class, behaves like the data generating one. Hence, such a universal process may be taken as a representation of the entire model class to be used for statistical inference. If $f(x^n)$ denotes the probability or density it assigns to the data string $x^n = x_1, \ldots, x_n$, then the negative logarithm $-\log f(x^n)$, which may be viewed as the shortest ideal code length for the data obtainable with the model class, is called the *stochastic complexity* of the string, relative to the considered model class. Unlike in related universal modeling, where the mean code length is sufficient, we also need an explicit asymptotic formula for the stochastic complexity. This is because it permits a comparison of different model classes by their stochastic complexity in accordance with the MDL (Minimum Description Length) principle.

The MDL principle for model selection and statistical inference in general is founded on the idea that the strength of the constraints in the data, imposed by the models, can be measured by the code length with which the data can be encoded, when advantage is taken of the constraints. This gives a data dependent criterion, which for its justification does not require the untenable assumption that the observed data are generated by some distribution. Hence, instead of minimizing a distance between the fitted models and the nonexisting 'true' distribution we just search for the model or model class that minimizes the code length.

The main problem in the implementation of the principle is how to estimate the shortest code length for the data, given a suggested model class. This can be difficult requiring ingenuity and hard work if the class of models is complex. Frequently a complex model class is built up of simpler ones, each fitted to a portion of the data, so that the total code length can be composed of the stochastic complexities of the components, and this again makes a formula for them useful. The seminal case is the class consisting of just one discrete distribution $P(x^n)$, for which the Shannon information $-\log P(x^n)$ may be taken to represent the shortest (ideal) code length among all prefix codes for a data sequence of a fixed length in the sense of the noiseless coding theorem; ie, in the mean.

The most important classes of models for which we can derive formulas for the stochastic complexity are of the type $M_k = \{f(x^n|\theta)\}$, or $M = \bigcup_k M_k$, each model indexed by a parameter vector $\theta = \theta_1, \ldots, \theta_k$ and satisfying the marginality condition for a random process. If the model class M_k is such that the maximum likelihood estimates satisfy the Central Limit Theorem for densities for each θ, an extention of the noiseless coding theorem states in broad terms that no process or, equivalently, code exists for which the mean lengths with respect to $f(x^n|\theta)$ for the various values of θ are shorter than the corresponding mean values of the stochastic complexity

$$-\ln f(x^n|\hat{\theta}(x^n)) + \frac{k}{2}\ln\frac{n}{2\pi} + \ln\int |I(\theta)|^{1/2}d\theta + o(1),$$

except for θ in a negligible subset. Here, $\hat{\theta}(x^n)$ denotes the maximum likelihood estimate, and $|I(\theta)|$ is the Fisher information. Moreover, this ideal code length up to terms of size o(1) is given by the negative logarithm of

$$f(x^n) = \frac{f(x^n|\hat{\theta}(x^n))}{\int f(x^n|\hat{\theta}(x^n))dx^n},$$

special cases of which were introduced and studied in [1] and [2]. The code length just given has the additional optimality properties for universal coding, [1], [3], that strengthen its distinguished status. The extension of the stochastic complexity formula to the larger family $M = \bigcup_k M_k$ is straightforward, and for a large subclass of finite alphabet Markov processes an efficient recursive implementation of the associated universal process is possible, [4]. However, the further extension to the case where the data generating processes are taken as nonparametric, residing in a suitable closure of the class M, poses difficulties. Some progress towards evaluating the redundancy has been made in [5] and [6]; for related results see also [7].

References

[1] L. D. Davisson, Minimax Noiseless Universal Coding for Markov Sources, *IEEE Trans. on Information Theory*, vol. IT-29, No. 2, pp. 211-215, March 1983.

[2] Yu. M. Shtarkov, Universal Sequential Coding of Single Messages, Translated from *Problems of Information Transmission*, vol. 23, No. 3, pp. 3-17, July-September 1987.

[3] B. S. Clarke and A. R. Barron, Information-Theoretic Asymptotics of Bayes Methods, *IEEE Trans. on Information Theory*, vol. IT-36, No. 3, pp. 453–471, May 1990.

[4] M. J. Weinberger, J. Rissanen, M. Feder, A Universal Finite Memory Source, submitted to *IEEE Trans. on Information Theory*

[5] P. Hall and E.J. Hannan, On Stochastic Complexity and Nonparametric Density Estimation, Biometrika, 75, 705-714, 1988.

[6] J. Rissanen, T.P. Speed, B. Yu, Density Estimation by Stochastic Complexity, *IEEE Trans. on Information Theory*, vol. IT-38, No. 2, pp. 315–323, March 1992.

[7] A.R. Barron, Y. Yang, B. Yu, Asymptotically Optimal Function Estimation by Minimum Complexity Criteria, *1994 IEEE International Symposium on Information Theory*, Trondheim, Norway, June 27-July 1, 1994.

Signal Expansions for Compression

Martin Vetterli[1]
Dept. of EECS, UC Berkeley, Berkeley CA 94720, USA

Abstract — **Signal expansions play a key role in practical compression schemes, from audio/image/video coding standards to current adaptives bases. Recent developments, especially related to wavelets series expansions, are reviewed, and current work on "best bases" is discussed.**

I. Introduction

Transform coding, together with predictive coding, is a key technique used in many practical compression systems. Its foundation is based on the Karhunen-Loève transform, which is the optimal transform under certain constraints [3]. In practice, approximations like the discrete cosine transform (DCT) are used, both for computational efficiency, and the fact that it is signal independent but still efficient for many practical signals. The transform coefficients are then quantized (usually in a scalar fashion) and entropy coded. This three block system [transform, quantization, entropy coding] raises some interesting questions:

- what are the best, possibly adaptive, transforms?
- what is the interplay of the three components?
- can successive approximation or multiresolution be efficiently achieved?

Recently, wavelets and their generalizations have appeared as alternatives to the more classic Fourier and DCT expansions [2]. In particular adapted expansions, and related algorithms to find the best bases, are an interesting extension.

II. Wavelet Series Expansions

Classically, windowed Fourier transforms have been used to obtain time-frequency representations of signals, and such representations are useful for source coding as well. Alternatively to local Fourier transforms, wavelet series have gained popularity. In this case, a particular prototype "mother" wavelet $\psi(t)$ is used to generate an orthonormal basis $\{\psi_{m,n}(t)\}$ for $L_2(R)$ by shifts and scales

$$\psi_{m,n}(t) = \frac{1}{2^{m/2}}\ \psi(t/2^m - n) \quad m,n \in Z.$$

A main difference between local Fourier expansions and wavelet series is that they provide a different tiling of the time-frequency plane. For example, at high frequencies or small scales, the wavelet is very sharp in time and acts like a mathematical microscope. In discrete-time, subband coding and filter banks permit the computation of sampled equivalents of local Fourier and wavelet transforms [7].

III. Adaptive Best Bases

Obviously, short-time Fourier transforms and wavelet series are only two out of a myriad of possible useful tilings. In particular, wavelet packets [1] and their time-varying generalizations [4] provide signal adaptive orthonormal bases. When the basis selection criterion is based on operational rate-distortion, we effectively have an adaptive transform coding algorithm, where quantization and entropy coding are included in the cost function. Since such transforms are adapted on the fly, computational efficiency is a must. In [4], a tree based pruning algorithm is used, while in [8] a dynamic programming procedure is applied.

IV. Overcomplete Representations

While orthonormal bases have many desirable properties as expansions for compression, a major drawback is their lack of shift-invariance. Overcomplete representations or frames overcome this problem, but the redundancy hurts compression. A recent result from oversampled analog to digital conversion [6] indicates that fine quantization in an orthonormal basis can be traded for coarse quantization in an overcomplete representation. Then, we discuss the use of matching pursuits [5] for compression applications. In matching pursuit, a very redundant dictionary is used together with a greedy algorithm to find a best approximation to a given signal. Choices of dictionaries and applications in video coding are considered.

V. Conclusion

A survey of signal expansions in the context of transform-type coding was given, with an emphasis of wavelets, adaptive and overcomplete representation. Expansions that adapt to the signals to be coded are a step towards *universal transforms* for compression.

Acknowledgements

The author would like to thank C.Herley, K.Ramchandran and N.T.Thao for their preprints and for useful discussions.

References

[1] R.R. Coifman and M.V. Wickerhauser, "Entropy-based algorithms for best basis selection," IEEE Trans. IT, Vol. 38, pp. 713-718, March 1992.

[2] I. Daubechies, **Ten Lectures on Wavelets**, SIAM, 1992.

[3] A.Gersho and R.M.Gray, **Vector Quantization and Signal Compression**, Kulwer Acad. Pub. 1992.

[4] C.Herley, J.Kovacevic, K.Ramchandran and M.Vetterli, "Tilings of the time-frequency plane: construction of arbitrary orthogonal bases and fast tiling algorithms." IEEE Trans. on SP, Vol. 41, pp. 3341-3359, Dec.1993.

[5] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," IEEE Transactions on SP, 3397–3415, Vol. 41, Dec. 1993.

[6] N.T.Thao and M.Vetterli, "Reduction of the MSE in R-times oversampled A/D conversion from $O(1/R)$ to $O(1/R^2)$," IEEE Trans. on SP, Vol. 42, No. 1, pp. 200-203, Jan. 1994.

[7] M.Vetterli and J.Kovacevic, **Wavelets and Subband Coding**, Prentice-Hall, 1995.

[8] Z. Xiong, C. Herley, K. Ramchandran and M. T. Orchard, "Flexible time segmentations for time-varying wavelet packets," Proc. Int. Symp. on TFTS, Philadelphia, Oct. 1994.

[1] This work was supported by grant NSF MIP-93-21302.

Large Deviations in Information Theory and Statistics

Amir Dembo

Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel

Abstract

Large deviations theory, a branch of probability theory that deals with estimates of probabilities of very rare events has close links with topics in information theory and in statistics. We shall explore some of these connections.

A sequence $\{\mu_n\}$ of Borel probability measures on $\mathcal{X}$ satisfies the large deviations principle (LDP) with a rate function $I : \mathcal{X} \to [0, \infty]$ if

UBD:

$$\limsup_{n\to\infty} \mu_n^{1/n}(F) \leq \exp(-\inf_{x\in F} I(x)) \quad \forall F \text{ closed}$$

LBD:

$$\liminf_{n\to\infty} \mu_n^{1/n}(G) \geq \exp(-\inf_{x\in G} I(x)) \quad \forall G \text{ open}$$

and $I^{-1}[0,\alpha]$ are compact sets for every $\alpha < \infty$. A sequence of r.v. is said to satisfy the LDP if their laws satisfy the LDP. The similarity of the LDP and the definition of weak convergence of probability measures is apparent.

Indeed, the theory of large deviations is soon reaching the state of maturity of weak convergence (for example, the texts [1,2,3] are dedicated to the former). However, from an application point of view these two theories serve complementary purposes. While weak convergence sheds light on the center of the distributions μ_n (i.e. events A for which $\mu_n(A)$ is bounded away from zero), large deviations theory deals primarily with the tails of μ_n.

Perhaps the most known LDP is Sanov's theorem, [1, Sec. 6.2], stating that for i.i.d. Σ-valued random variables $\{X_i\}_i$ each distributed according to μ, their empirical measures $L_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ satisfy in $M_1(\Sigma)$ (= space of Borel probability measures on Σ) the LDP with rate function $H(\cdot|\mu)$. Here $H(\cdot|\cdot)$ stands for the relative entropy (also known as the KL divergence or cross entropy),

$$H(\nu|\mu) = \begin{cases} \int_\Sigma f \log f d\mu & \frac{d\nu}{d\mu} = f, \text{ exists} \\ \infty & \text{otherwise} \end{cases}$$

Sanov's theorem provides a clue to some of the links explored in this talk, namely:

(a) The method of types, introduced in Information Theory (c.f. [4,5]), allows one to prove Sanov's theorem when Σ is a finite set and goes much beyond this simple setup.

(b) The decisive role of the relative entropy in Sanov's theorem is exemplified in the use of information inequalities to prove statements about conditional laws (c.f. [6]). The notions of sufficient statistics and of universal prior are closely related to these conditional laws.

(c) Sanov's theorem yields the asymptotics of probability of error in the Hypothesis Testing problem (c.f. [1, Sec. 3.4]).

(d) The map $\nu \mapsto \int x d\nu : M_1(\Sigma) \to \Sigma$ contracts Sanov's theorem to Cramer's theorem dealing with the LDP for the empirical means $\hat{S}_n = \frac{1}{n}\sum_{i=1}^n X_i$. The corresponding result for weakly dependent X_i (Gärtner-Ellis theorem) provides a large-deviations-based proof of Shannon's (noisy) source coding theorem (c.f. [1, Sec. 3.6]).

Time permitting, other relations to be discussed are the use and value of large deviations in non-parametric and/or sequential statistics problems, in certain (practical) communication theory problems (c.f. [7]), and in the study of fractal measures and sets.

References

[1] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett, Boston, 1993.

[2] J. D. Deuschel and D. W. Stroock. *Large Deviations.* Academic Press, Boston, 1989.

[3] S. R. S. Varadhan. *Large Deviations and Applications.* SIAM, Philadelphia, 1984.

[4] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic Press, New York, 1981.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

[6] I. Csiszàr. Sanov property, generalized I–projection and a conditional limit theorem. *Ann. Probab.*, 12:768–793, 1984.

[7] J. A. Bucklew. *Large Deviations Techniques in Decision, Simulation, and Estimation.* Wiley, New York, 1990.

The Entropy Strategy for Shape Recognition

Donald Geman[1]

Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003, USA

Abstract —

We consider a computational strategy for shape recognition based on choosing "tests" one at a time in order to remove as much uncertainty as possible about the true hypothesis. The approach is compared with other recognition paradigms in computer vision and illustrated by attempting to classify handwritten digits and track roads from satellite images.

I. Overview

We explore the possibility of recognizing shapes "simply" be asking the right questions in the right order. In the abstract formulation, we are given a finite list of possible "hypotheses"; exactly one of these is true and we wish to decide which it is based on the results of various "tests" or "questions." There is a decision tree which instructs us how to perform the tests and eventually classify the results. Each interior node of the tree is assigned one of the tests and each terminal node is assigned one of the hypotheses. The assignment of tests, the "strategy," is adaptive in the sense that the choice of the test at each node may depend on the test values observed at all preceding nodes along the same branch. Ideally, the choice would be driven by some global measure of efficiency, such as achieving the most accurate classifier for a given average number of tests, or reaching the fastest decision at a given level of accuracy. But these problems are intractable, and we shall opt instead for the "greedy" strategy in which the tests are chosen recursively based on minimizing the expected entropy of the updated disribution over hypotheses given the test results.

We have applied this to two problems in shape recognition, focusing on linear, deformable structures. The raw data is a binary or grey-level image, the tests are particular "features" (i.e., image functionals), and the hypotheses refer to particular shape classes or spatial positionings.

II. Roads

This application is joint work with Bruno Jedynak of INRIA- Rocquencourt. We describe a new algorithm [2] for tracking major roads from panchromatic SPOT satellite imagery, demonstrated on SPOT images of size 6000×8000, representing a $60km \times 80km$ square on the ground, in this case in southern France.

The standard construction of decision trees (e.g., in coding, CART [1], and machine learning) is off-line, nonparametric, and based on "training data." However, in our formulation of tracking, it is impossible to pre-compute and store the entire decision tree: it has too many branches from each (interior) node, it is too deep (i.e., too many tests are needed to reach a decision), and the number of possible road locations is enormous. (The tests are local matched filters indexed by image location and designed to respond to short road segments.) Instead, the entropy strategy is implemented *on-line*: each new filter is chosen during the actual tracking based on the particular filter results previously encountered; in other words, we only compute the branch of the tree that is needed for the data at hand. In fact, the amount of time necessary to perform the tests is small compared to determining the "right" test to perform. On the other hand, compared to maximum likelihood estimation, the number of tests actually performed until a decision is made is exponentially small; indeed, maximum likelihood is computationally impossible.

Our approach is also model-based rather than nonparametric. As a result, we can formulate the problem of minimizing entropy in explicit and relatively simple analytical terms. To execute the strategy we then alternate between data collection and optimization: at each iteration, new image data is examined and a new entropy minimization problem is solved (exactly) resulting in a new image location to inspect, and so forth. This will be illustrated with a video.

III. Digits

We shall also briefly mention another application - the recognition of handwritten numerals - which is co-authored with Prof. Yali Amit of the University of Chicago. There are ten hypotheses and the strategy is again constructed by stepwise entropy reduction, but *off-line* and not in the standard Euclidean framework. Instead, we construct *relational* classification trees based on accumulating information about a *graphical representation* of the image data involving planar arrangements among local landmarks. Actually, we construct many, each being a distribution-valued test. For any given training set, there is then a fundamental trade-off between the tree *depth* and tree *generality*; this is related to well-known issues and tradeoffs in computational learning and computer vision. Finally, the classification rates obtained appear to be comparable to state-of-the-art neural networks and other nonparametric statistical classifiers.

Acknowledgements

The road detection work was carried out at Project Syntim at INRIA- Rocquencourt; we are grateful to Andre Gagalowicz and Jean Philippe Roze for their valuable support.

References

[1] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA., 1984

[2] D. Geman and B. Jedynak, "Tracking Roads in Satellite Images by Playing Twenty Questions," preprint, 1994

[1]This work was supported by NSF Grant DMS-9217655 and ONR Contract N00014-91-J-1021

SESSION I

Stochastic Complexity and Universal Data Compression

Maximum entropy and related methods

Imre Csiszár[1]
Mathematical Institute of the Hungarian Academy of Sciences
H–1364 Budapest, P.O.B. 127, Hungary

Originally coming from physics, maximum entropy (ME) has been promoted to a general principle of inference primarily by the works of Jaynes [4].

ME applies to the problem of inferring a probability mass (or density) function, or any non-negative function $p(x)$, when the available information specifies a set E of feasible functions, and there is a prior guess $q \notin E$. The ME solution is that $p^* \in E$ which minimizes the information divergence

$$D(p \parallel q) = \sum [p(x) \log \frac{p(x)}{q(x)} - p(x) + q(x)]. \quad (1)$$

For probability mass functions, if q is uniform, minimizing (1) is the same as maximizing the entropy $H(p)$. This is why the method is called ME also in general.

In typical applications, the available information consists in linear constraints on p, i.e.,

$$E = \{p: \sum p(x) a_i(x) = b_i \ , \quad i = 0, 1, \ldots, k\}. \quad (2)$$

Then the ME solution p^* (uniquely) exists, and

$$p^*(x) = q(x) \exp \sum_{i=0}^{k} \vartheta_i \ a_i(x) \ , \quad (3)$$

providing q is strictly convex and E contains any strictly positive p. In the non-discrete case (with sums replaced by integrals), the existence of ME solution can not be asserted in the above generality, although a unique p^* always exists, possibly not in E, such that $D(p_n \parallel p^*) \to 0$ for every $\{p_n\} \subset E$ with $D(p_n \parallel q) \to \inf_{p \in E} D(p \parallel q)$; this p^* is of form (3), cf. [1].

We will review the arguments that have been put forward for justifying ME. In this author's opinion, the strongest theoretical support to ME is provided by the axiomatic approach. This shows that, in some sense, ME is the only logically consistent method of inferring a function subject to linear constraints. This approach also leads to alternatives that come into account under weaker axioms, cf. [2]. Such are the methods of minimizing an f-divergence

$$D_f(p \parallel q) = \sum q(x) f(\frac{p(x)}{q(x)}) \quad (4)$$

or a Bregman divergence

$$B_f(p \parallel q) = \sum [f(p(x)) - f(q(x)) - f'(q(x))(p(x) - q(x))], \quad (5)$$

where f is a strictly convex function. Minimizing (5) leads to scale invariant inference if $f = f_\alpha$ where $f_1(t) = t \log t - t$, $f_0(t) = -\log t$, $f_\alpha(t) = t^\alpha$ if $\alpha > 1$ or $\alpha < 0$, $f_\alpha(t) = -t^\alpha$ if $0 < \alpha < 1$. Inference by minimizing (5), particularly with $f = f_\alpha$, has been suggested also in [5], based on another axiomatic approach.

For the problem of attainment of the minimum of (4) or (5) subject to $p \in E$ (in the non-discrete case), there is an analogue of the result in the passage containing (3). It depends on the behavior of f at infinity whether or not this permits to give simple sufficient conditions for the minimum to be attained, cf. [3].

[1]This work was supported by OTKA Grant No.1906

References

[1] I. Csiszár, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probab.*, vol. 12, pp. 768-793, 1984.

[2] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, vol. 19, pp. 2031-2066, 1991.

[3] I. Csiszár, "Generalized projections for positive valued functions," *Acta Sci. Math. Hungar.*, to appear.

[4] E. T. Jaynes, *Papers on Probability, Statistics, and Statistical Physics*, R. D. Rosenkrantz, ed., Reidel, Dordrecht, 1983.

[5] L. K. Jones and C. L. Byrne, "General entropy criteria for inverse problems," *IEEE Trans. Inform. Theory*, vol. 36, pp. 23-30, 1990.

A Stronger Version of the Redundancy-Capacity Theorem of Universal Coding

Neri Merhav and Meir Feder

Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, ISRAEL
Department of Electrical Engineering - Systems, Tel Aviv University, Tel Aviv 69978, ISRAEL

Abstract — **The capacity of the channel induced by a given class of sources is well known to be an attainable lower bound on the redundancy of universal codes w.r.t this class, both in the minimax sense and in the Bayesian (maximin) sense. We show that this capacity is essentially a lower bound also in a stronger sense, that is, for "most" sources in the class. This result extends Rissanen's lower bound for parametric families. We demonstrate its applicability in several examples and discuss its implications in statistical inference.**

In universal coding w.r.t a given class of sources the objective is to design a single code that "performs well" for every source in the class. The sources in the class are indexed by a variable $\theta \in \Lambda$. The performance of a given code L, is judged on the basis of the redundancy which is defined as the difference between the expected code length of L w.r.t a given source P_θ and the nth order entropy of P_θ, normalized by the length n of the input vector.

Two important notions of universality are the *maximin* universality and the *minimax* universality [1]. Gallager [2] was the first to show that the minimax redundancy and the maximin redundancy are equivalent and that they are both equal to the capacity of the "channel" whose input is θ and whose output is the random source vector $X^n = (X_1, ..., X_n)$. In particular, for parametric families where θ is a k-dimensional vector, the minimax redundancy, and hence also the maximin redundancy and the capacity of the corresponding channel, was shown to be essentially $0.5k \log n/n$.

Rissanen [3] has strengthened the notion of universality w.r.t parametric families by showing that $0.5k \log n/n$ is not only an achievable lower bound in the minimax sense, but also a lower bound for "most" sources in the class. Here by "most" sources we mean *every* point θ except for a subset of points whose Lebesgue measure vanishes as n grows. Rissanen's proof, however, relies heavily on the structure of the parametric family and essentially the main insight that can be gained from his work is that the redundancy is strongly related to the richness of the class, which in the parametric case is proportional to the dimension k of the parameter vector.

It turns out thar Rissanen's stronger notion of universality extends to the general case where the class of sources is not necessarily a parametric family. Specifically, we show that the Shannon capacity of the induced channel is a lower bound on the redundancy that holds simultaneously for all sources in the class except for a subset of points whose probability, under the capacity-achieving probability measure, is vanishing as n tends to infinity. This means that the minimax redundancy and the lower bound essentially coincide for most choices of θ. Moreover, if the capacity-achieving probability density happens to be positive almost everywhere (Lebesgue), as is normally the case in parametric families, the above result holds also for most sources in the Lebesgue measure sense and therefore Rissanen's result is obtained as a special case.

The proof is completely different and considerably simpler than Rissanen's proof [3]. However, it does not allow a free choice of any prior, other than the capacity-achieving prior, that might be reasonable as well for weighting the set of points that violate the bound. We next provide another variant of our result which permits any prior on the index set, but then the *random coding* capacity of the induced channel rather than its Shannon capacity is obtained as a lower bound. Here the random coding capacity refers to the normalized logarithm of the maximum number M of randomly chosen points $\theta_1, ..., \theta_M$, which form, with high probability, a set of distinguishable sources $P_{\theta_1}, ..., P_{\theta_M}$. For most cases of practical interest the Shannon capacity and the random coding capacity are equivalent and hence the resulting bound is virtually as tight. We believe that another advantage of this random coding capacity result is that it may add some new insight about the relation between redundancy and capacity. Specifically, in the proof of this result the redundancy is linked directly, not only to the mathematical notion of capacity as the maximum mutual information, but also to the *operational* notion, i.e., the maximum achievable rate of reliable communication.

The results above have a broader significance in statistical inference. In the absence of knowledge about the true underlying class member P_θ, the statistician wishes to construct a single *universal* probability measure Q that "explains well" the data. His task is successful if Q is simultaneously "close" to every source in the class , where distance is measured in terms of the divergenece $D(P_\theta||Q)$. In this context, our main result is the following attainable lower bound: For all $\theta \in \Lambda$, except for points in a subset $B \subset \Lambda$ whose probability, under the capacity-achieving prior, vanishes as $n \to \infty$,

$$D(P_\theta||Q) \triangleq E_\theta \log \frac{P_\theta(X^n)}{Q(X^n)} \geq (1-\epsilon)C_n$$

where C_n is the capacity of the channel from θ to X^n. Thus, a necessary and sufficient condition that the statistician task will asymptotcally succeed is that $C_n/n \to 0$. Note that, unlike the lossless data compression problem, this setting applies to the continuous alphabet case as well. This point of view provides a general framework for the choice of a statistical model in the presence of uncertainty, which can be used for other decision making problems as well, like universal gambling, portfolio selection, and prediction.

References

[1] L. D. Davisson, "Universal Noiseless Coding," *IEEE Trans. Inform. Theory*, Vol. IT-19, No. 6, pp. 783–795, November 1973.

[2] R. G. Gallager, "Source Coding with Side Information and Universal Coding," unpublished manuscript, Sept. 1976.

[3] J. Rissanen, "Universal Coding, Information, Prediction, and Estimation," *IEEE Trans. Inform. Theory*, Vol. IT-30, No. 4, pp. 629–636, July 1984.

Bounds on Universal Coding: The Next Generation

Jacob Ziv

(Joint work with Yehuda Hershkovits)

Department of Electrical Engineering, Technion, Haifa 32000, ISRAEL

Abstract — **An important class of universal encoders is the one where the encoder is fed by two inputs: a) The incoming string of data to be compressed. b) An N-bit description of the source statistics (i.e. a "training sequence"). We consider Fixed-to-Variable universal encoders that noiselessly compress blocks of length ℓ.**

Two problems will be addressed:

1. The Minimum Training-sequence length, $N_{min}(\ell)$:
   Given a class of admissible stationary sources, find the minimal length of a training sequence needed in order to guarantee that any source in the given class, with an l-th order entropy $H_\ell \leq$ logA, will yield some compression (A is the alphabet size).
2. An Optimal Universal Encoder (UE):
   Find a UE that "ensures" that the compression for EVERY source in the given class is close to the minimal possible compression H_ℓ, once the training sequence is longer than $N_{min}(\ell)$.

The first case to be considered is the one where the training sequence and the incoming data string are assumed to be statistically independent.

A Converse Theorem (solving problem (1)) and a Coding Theorem (solving problem (2)) are given for the class of finite-alphabet stationary sources with a vanishing memory(i.e. sources that satisfy a certain mixing condition [1], [3]).This class includes all finite-order Markov sources.

Another, perhaps more practical case is the one where the training sequence consists of the last N bits of the data that has been processed.(i.e. a "sliding window" algorithm).

For any stationary source P over an alphabet of A letters, let $B_N = [X_{-j}^0 : j = max[i : P(X_{-i}^0 \geq 1/N); i = -1, 0, 1, 2, ...]]$ and define the conditional entropy $H_\ell^N(X_1^\ell | X_{-j}^0)$ which is monotonically decreasing with N , and satisfies $H \leq H_\ell^N(X_1^\ell | X_{-j}^0) \leq H_\ell$. It is demonstrated that (for large N) the length of the training must be bigger than N, or else, for any universal FV encoder for ℓ-vectors there exists at least one stationary source with $H_\ell^N(X_1^\ell | X_{-j}^0) \leq R \leq$ logA for which the compression is logA-ϵ . Here $\epsilon \geq 0$ and $0 \leq R \leq$ logA are arbitrary and ℓ, the length of the source words, must be of order between loglogN and logN.Conversely, we describe a compression scheme that yields a compression that is arbitrarily close to $H_\ell^N(X_1^\ell | X_{-j}^0)$ for every stationary source, provided that the length of the sequence is larger than N (i.e. the length is at least $N^{1+\epsilon}$ where ϵ is arbitrarily small).

The coding theorems are based on variants of the Lempel-Ziv data-compression algorithm [2].

References

[1] Wyner,A.D. and J.Ziv, "Some asymptotic properties of a stationary ergodic source with application to data compression", IEEE Transactions on Information Theory, Vol. IT-36, November 1989, pp. 1250-1258.

[2] [2] Ziv, J. and A.Lempel, "A universal algorithm for sequential data compression", IEEE Transactions on Information Theory, Vol. IT–24, May 1977. pp. 337–343.

[3] Wyner, A. D, and J.Ziv, "Classification with Finite-Memory", submitted to the IEEE Transactions on Information Theory.

Jeffreys' Prior Yields the Asymptotic Minimax Redundancy

Bertrand S. Clarke and Andrew R. Barron

Stat. Dept., U.B.C., 2021 West Mall, Vancouver, Canada, V6T 1Z2 and Stat. Dept., Yale Univ. PO Box 208290, Yale Station, New Haven, CT 06520, USA

Abstract — **We determine the asymptotic minimax redundancy of universal data compression in a parametric setting and show that it corresponds to the use of Jeffreys prior. Statistically, this formulation of the coding problem can be interpreted in a prior selection context and in an estimation context.**

I. Introduction

Here we exploit a relationship between coding in information theory and risk in statistics. In source coding one often wants to minimize the redundancy of the code and in channel coding one often wants to achieve a high rate of transmission. These two goals can be defined in terms of the relative entropy between two distributions. The relative entropy between distributions can also be used as a loss function in a decision theory context. Since, roughly speaking, a source code corresponds to an estimator for an unknown distribution the statistical implication is that one can seek Bayes estimators and a least favorable prior which has desired noninformativity properties.

II. Main results

If one has data from a source distribution with density given by $p_\theta(x^n) = \Pi_{i=1}^n p_\theta(x_i)$ where p_θ is a member of a smooth parametric family and one has a continuous density w on the parameter space then the Bayes code achieves the minimum of the Bayes redundancies $\int w(\theta) D(p_\theta^n || q_n) d\theta$ over all distributions q_n, where D is the relative entropy. The Bayes redundancy for a code is the same as the Bayes risk for the estimator corresponding to that code. Thus, the Bayes code is based on the mixture density $m(x^n) = \int w(\theta) p_\theta(x^n) d\theta$. Equivalently, m can be regarded as a Bayes estimator, with risk $D(p_\theta^n || m_n)$, and Bayes risk $\int w(\theta) D(p_\theta^n || m_n) d\theta$. Maximizing the Bayes redundancy over choices of w gives the maximin redundancy, or maximin risk. One can therefore identify a maximin estimator or a maximin code, formed from the mixture distribution with respect to the choice of w achieving the maximal Bayes redundancy.

Alternatively, one might seek a code or estimator which minimizes the worst case redundancy. Game theory suggests that such a minimax procedure will be the same as the maximin procedure. This turns out to be the case and one can identify a least favorable prior w.

Our main results, see [3], are as follows. First, the redundancy of the Bayes code i.e., the risk of the Bayes estimator m_n is

$$R_n(\theta, w) = \frac{d}{2} \log \frac{n}{2\pi e} + \log \frac{\sqrt{|I(\theta)|}}{w(\theta)} + o(1),$$

uniformly for θ in compact sets K in the interior of the parameter space, where $I(\theta)$ is the Fisher information matrix. Second, the Bayes redundancy of the Bayes code, i.e., the Bayes risk of the Bayes estimator is

$$R_n(w) = \frac{d}{2} \log \frac{n}{2\pi e} + \int_K w(\theta) \log \frac{\sqrt{|I(\theta)|}}{w(\theta)} d\theta + o(1).$$

Third the asymptotic minimax redundancy, i.e., the minimax risk is

$$R_n = \frac{d}{2} \log \frac{n}{2\pi e} + \log \int_K \sqrt{|I(\theta)|} d\theta + o(1) = R_n^*,$$

where R_n^* is the maximin redundancy, or the maximin risk. Finally, the least favorable prior is seen to be Jeffreys prior given by $|I(\theta)|^{1/2} / \int_K |I(\theta)|^{1/2} d\theta$, see [5]. With this choice the redundancy, or risk, achieves asymptotically the same value $R_n(\theta, w) = R_n$, uniformly for θ in compacta interior to K.

The associated codelength takes the form

$$\log \frac{1}{m(x^n)} = \log \frac{1}{p_{\hat{\theta}}(x^n)} + \frac{d}{2} \log \frac{n}{2\pi} + \log \int_K \sqrt{|I(\theta)|} d\theta + o(1)$$

where $\hat{\theta}$ is the maximum likelihood estimator see [4]. This mixture codelength has been suggested for use in a suitable formulation of the minimum description length principle for model selection, see [1], [6].

III. Implications

This least favorable prior has an interpretation in channel coding. The Bayes risk $R_n(w)$ is the Shannon mutual information $I(\Theta; X^n)$. This corresponds to the channel in which θ is sent to n receivers who decode the message x^n together. Asymptotically, the source achieving the channel capacity, $\max_w I(\Theta; X^n)$, is Jeffreys' prior.

In statistics, the distribution achieving the maximal mutual information is called a reference prior. The results for continuous parameters provide formal verification of a conjecture in [2], that in the absence of nuisance parameters, reference priors are Jeffreys priors. Equivalently, one can write the mutual information as $E_m D(w(\cdot|X^n) || w(\cdot))$. The w maximizing this quantity asymptotically gives a posterior as far as possible from the prior on average. That is, it is the prior which leaves the most to be learned from the data and so represents minimal informativity.

References

[1] A. R. Barron, "Logically smooth density estimation," Ph. D. thesis, Stanford University, 1985.

[2] J. M. Bernardo, "Reference posterior distributions for Bayesian inference," *J. Roy. Statist. Soc. Ser. B*, Vol. 41, pp. 113-147, 1979.

[3] B. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Statist. Planning and Inference*, vol.41, 37-60, 1994.

[4] B. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, 453-471, 1990.

[5] H. Jeffreys, *Theory of Probability*. New York: Oxford University Press, 1961.

[6] J. Rissanen, "Fisher information and stochastic complexity," to be presented at 1994 IEEE-IMS Workshop.

Lower Bounds on Expected Redundancy

Bin Yu

Department of Statistics, University of California, Berkeley, CA 94720-3860, USA

Abstract — **This paper focuses on lower bound results on expected redundancy for universal compression of iid data from parametric and nonparametric families. Two types of lower bounds are reviewed. One is Rissanen's almost pointwise lower bound and its extension to the nonparametric case. The other is minimax lower bounds, for which a new proof is given in the nonparametric case.**

I. Introduction

One important ingredient of Rissanen's Stochastic Complexity theory is his (almost) pointwise lower bound on expected redundancy for regular parametric models (cf. [4]), and a minimax counterpart follows from [2]. By expressing expected redundancy in terms of accumulated expected prediction errors, a similar lower bound was proved in [5] and [7] on expected redundancy for a smooth nonparametric class of densities. This lower bound was shown in two different senses: one extending the parametric pointwise bound to an "artificial" parameter space with a dimension depending on the sample size ([5]), and the other in the minimax sense ([7]). In this paper, we review these lower bounds and the methods used to prove these lower bounds. Finally we provide a new proof for the lower bound in the nonparametric case. This new proof is information-theoretic, bypassing the detour to accumulated prediction errors, although we do borrow calculations from the density estimation literature.

II. Rissanen's Lower Bound and its Nonparametric Extension

For a given iid data string $x_1, x_2, ..., x_n$ and without knowing the distribution f which generated the data, we would like to compress the data in an efficient way. When $f(x) = f(x|\theta)$ belongs to a smooth k dimensional parametric family such that the parameter θ can be estimated at the $n^{-1/2}$ rate, Rissanen [4] showed that we need at least $H(f) + \frac{k}{2}\frac{\log n}{n}$ bits per data point, asymptotically. This lower bound holds in expectation and it holds for almost all parameter values in the parameter space. With a prefix code achieving this lower bound, Rissanen justified that $\frac{k}{2}\frac{\log n}{n}$ can be viewed as the coding complexity measure of the model class.

When f is known to be in the smooth nonparametric density class of bounded derivatives on [0,1], a complexity rate measure of $n^{-2/3}$ was established in [5] by embedding the nonparametric class in a parametric class of dimension of order $n^{1/3}/\log n$. This embedding reflects the fact that a smooth nonparametric class is in essence a parametric class.

III. The Minimax Lower Bounds

Through the minimax theorem (cf. [3]), the minimax expected redundancy over a parametric class is equivalent to the maximum of the Bayes redundancy which is the same as the mutual information. Fortunately, for a given prior, the Bayes code is the mixture density with respect to that prior and the first term in the expansion of the Bayes redundancy or mutual information is obtained in [2] to be the Rissanen coding complexity of $\frac{k}{2}\frac{\log n}{n}$. Hence this complexity measure also serves as the minimax lower bound on expected redundancy.

For the nonparametric class mentioned above, the minimax theorem still holds; therefore any Bayes redundancy or mutual information provides a lower bound. However, no prior seems to exist on the whole density class for which the Bayes redundancy can be approximated analytically. On the other hand, the expected redundancy is simply the accumulated prediction or estimation error in terms of Kullback-Leibler divergence, and techniques have long been developed to obtain minimax lower bounds on density estimation errors in the nonparametric case, cf [6]. By lower bounding the divergence by the Hellinger distance and borrowing Assouad's technique, a minimax rate lower bound of $n^{-2/3}$ was established in [7].

Note that in applying Assouad's technique, one does not calculate the Bayes estimation error over the whole class, but only over a conveniently chosen hypercube sub-class, and the Bayes estimation error over this sub-class provides a lower bound on the minimax estimation error. It turns out that this detour to accumulated prediction or estimation error is not necessary since we can use the hypercube sub-class directly with the redundancy. Using a result from the density estimation literature ([1]), it can be shown that $n^{-2/3}$ gives the rate in a lower bound. Thus we obtain a new proof for the minimax rate lower bound in [7], and this new line of proof applies to other smooth classes of densities.

Superficially, the proof in the parametric case has a continuous flavor since it relies on nice continuous priors on the whole parameter space, whereas the proof in the nonparametric case has a discrete flavor because of the hypercube sub-class it relies on. In essence, however, the former is also discrete since the continuous prior can be replaced by a discrete uniform prior sitting on a grid sub-set of the parametric space, as long as the grid-size is of the order or smaller than $n^{-1/2}$. Note that the nearest neighbors on the hypercube also have Hellinger distances of order $n^{-1/2}$–the rate at which n iid data points can possibly distinguish two distributions.

References

[1] L, Birgé and P. Massart, "Estimation of integral functionals of a density," MSRI Tech. Report 024-92, 1992.

[2] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. on Information Theory*, **36** 453-471, May, 1990.

[3] I. Csiscár, "Information theoretical methods in statistics," *Class notes*, University of Maryland, College Park, MD, Spring, 1990.

[4] J. Rissanen, "Stochastic complexity and modeling," *Annals of Statistics*, **14** 1080-1100, 1986.

[5] J. Rissanen, T. Speed and B. Yu, "Density estimation by stochastic complexity," *IEEE Trans. on Information Theory*, **38** 315-323, May, 1993.

[6] B. Yu, "Assouad, Fano, and Le Cam," To appear in *Festschrift in Honor of L. Le Cam on His 70th Birthday*, 1995.

[7] B. Yu and T. Speed, " Data compression and histograms," *Probability Theory and Related Fields*, **92** 195-229, 1992.

When is the weak rate equal to the strong rate?

Paul C. Shields[1]

Mathematics Departments, U. of Toledo, Toledo, OH 43606, and Eötvös Lorand University, Budapest.

Abstract — **A condition on a class of processes guaranteeing that the weak redundancy rate has the same asypmtotic order of magnitude as the strong redundancy rate will be discussed.**

I. Introduction

In recent papers, examples were constructed showing that there is no nice weak redundancy rate for various subclasses of the class of ergodic processes, [1, 2, 3]. In each of these it was shown how to find a process in the class whose n-th order reduncancy was large, then it was shown how to make small changes to produce a process whose m-th order redundancy was large, for some $m > n$. A suitable passage to a limit then produced the desired example. The two essential features were the existence of processes with large redundancy in a neighborhood of any member of the class and a completeness property to insurance the existence of a limit. The purpose of this talk is to formalize these two features, in order to clarify the prior constructions and extend them.

II. Notation and terminology.

The (expected) redundancy of a prefix n-code C_n, relative to a process P, is defined by

$$R(C_n|P) = E(L_n|P) - H_n(P),$$

where $E(L_n|P)$ is expected code length and

$$H_n(P) = -\sum_{a_1^n} P(a_1^n)\log P(a_1^n),$$

is the n-th order entropy of P. The minimax expected redundancy for a class $\mathcal{S}$ of stationary processes with alphabet A is defined by

$$\bar{R}_n(\mathcal{S}) = \min_{C_n}\max_{P\in\mathcal{S}} \bar{R}(C_n|P),$$

where the minimum is over all binary prefix n-codes.

A sequence $\{C_n\}$ is called a prefix-code sequence if C_n is a binary prefix n-code, for each n. A nondecreasing function $n \mapsto \rho(n)$ is called a *strong rate* for the class $\mathcal{S}$ if there is a constant M such that $\bar{R}_n(\mathcal{S}) \le M\rho(n)$, $n \ge 1$, and if, for any prefix-code sequence $\{C_n\}$ and any function $\phi(n) = o(\rho(n))$, there is a member $P \in \mathcal{S}$ such that $R(C_n|P)/\phi(n)$ is unbounded.

A nondecreasing function $n \mapsto \rho(n)$ is called a *weak rate* for the prefix-code sequence $\{C_n\}$ on the class $\mathcal{S}$ if for each $P \in \mathcal{S}$ there is a finite number $M = M(P)$ such that

$$R_n(C_n|\mu) \le M\rho(n),\ n \ge 1, \qquad (1)$$

and if, for any prefix-code sequence $\{C_n\}$ and any function $\phi(n) = o(\rho(n))$, there is a member $P \in \mathcal{S}$ such that $R(C_n|P)/\phi(n)$ is unbounded. (Note that weak rates allow the constant M to depend on P.)

[1]Partially supported by NSF grant DMS-9024240 and MTA-NSF project 37.

Let $d(P,Q)$ be a metric on a class $\mathcal{S}$ of stationary processes, and let $N_\epsilon(P) = \{Q \in \mathcal{S}\colon d(P,Q) < \epsilon\}$ denote the ϵ neighborhood of P. The metric space $(\mathcal{S}, d)$ is called *locally rich* if $N_\epsilon(P)$ and $\mathcal{S}$ have the same strong rate, for every $P \in \mathcal{S}$ and $\epsilon > 0$. The following theorem will be proved.

III. The weak-rate/strong-rate theorem.

If $\mathcal{S}$ has a locally rich, complete metric d such that d-convergence implies weak convergence and convergence in entropy, then the weak rate and strong rates for $\mathcal{S}$ are the same.

Proof: Select $P^{n(i)} \in \mathcal{S}$ such that $d(P^{n(i+1)}, P^{n(i)}) \le \delta_i$, and such that $R(C_{n(i)}|P^{n(i)}) > M_j\phi(n(j))$, $j \le i$, where $M_i \to \infty$, and $\sum_i \delta_i < \infty$.

IV. Summary table.

Class	Strong rate	Weak rate
i.i.d.	$\log n$	$\log n$
Markov (k)	$\log n$	$\log n$
Finite state (M)	$\log n$	$\log n$
Renewal	$n^{1/2}$	$n^{1/2}$
M-Renewal (k)	$n^{(k+1)/(k+2)}$	$n^{(k+1)/(k+2)}$
Regenerative	n	n
B-processes	n	n
Ergodic processes	n	n
$\cup_k$ Markov (k)	n	$\log n$
Finite state	n	$\log n$
Renewal/finite	$n^{1/2}$	$\log n$
M-Renewal (k)/finite	$n^{(k+1)/(k+2)}$	$\log n$
Regenerative/finite	n	$\log n$
$\cup_k$ M-Renewal (k)	n	$n/\log n^*$

Class/ finite = finitely many waiting times.
*-upper bound only.

Metrics:

1. Markov (k): $(k+1)$-order variational distance.
2. Renewal: $d(P,Q) = \sum t|P(t) - Q(t)|$, where $P(t) =$ probability that waiting time is t.
3. Regenerative:
   $D(P,Q) = d(P,Q) + \sum_t \sum_{a_1^t} |P_t(a_1^t)P(t) - Q_t(a_1^t)Q(t)|$,
   where $P_t = t$-order distribution given that waiting time is t, and $d(P,Q)$ is the renewal distance.
4. B-processes: $\bar{d}$-metric.
5. Ergodic processes: $\bar{f}$-metric.

References

[1] P. Shields, "Universal redundancy rates do not exist," IEEE Trans. Inform. Th., IT-39(1993), 520-524.

[2] P. Shields and B. Weiss, "Universal redundancy rates for B-processes do not exist," IEEE Trans. Inform. Th., to appear.

[3] "Redundancy rates for renewal and other processes," IEEE Trans. on Inform. Th., to be submitted.

SESSION II

Vector Quantization, Classification and Regression Trees

Variable-rate, Lossy, Tree-structured Codes and Digital Radiography

Richard A. Olshen
Division of Biostatistics, Stanford University School of Medicine, Stanford, California 94305–5092
olshen@playfair.stanford.edu

Abstract — **My talk is a survey of binary tree-structured methods for clustering as they apply to predictive, pruned, tree-structured vector quantization (predictive PTSVQ). Much of the material concerns applications of PTSVQ to the lossy coding of digital medical images, especially CT and MR chest scans. There is a brief introduction to the asymptotic properties of the algorithms and to the attempt to understand variability and covariability of amino acids in the V3 loop region of HIV. The research has been collaborative with many others over a five year period.**

The algorithms involve successively partitioning the range of a set of pixel vectors $\mathbf{X}$, and can be viewed as successive two-means clustering. When $\mathbf{X} \in \mathcal{R}^k$ the partitioning is by hyperplanes; qualitative data can be handled, too, as in the application to HIV amino acid sequences. Results are summarized by a binary tree; a pixel vector $\mathbf{X}^*$ to be coded is passed from the root node successively to a terminal node (t). The codeword assigned is simply a suitably defined centroid of learning sample $\mathbf{X}$ values at t. The *bit rate* is the average depth of the tree. Splitting is always "greedy," in senses to be described. We grow an initial tree larger than we intend to use and prune it back to smaller ones. Every subtree of the large initial tree has its own figure of merit and assigned "penalty" for complexity. For a given penalty, there is a unique smallest pruned subtree of the cited initial large tree that is optimal in terms of figure of merit. As the penalty increases, the sequence of optimally pruned subtrees is nested. Codes that correspond to the sequence of optimally pruned subtrees are thus seen to have a natural progressive property.

Versions of these algorithms can be shown to be "consistent" [4] in a sense to be described. Part of the argument involves showing that the algorithm terminates when it is applied with a bit rate constraint to a fixed absolutely continuous distribution with compact support. Next, a continuity property is established relative to a fixed, convergent sequence of distributions. Finally, aspects of empirical processes are brought to bear upon the large sample behavior of the algorithm when the learning sample is, beyond the cited assumptions, stationary and ergodic.

Applications of PTSVQ to problems of data compression in digital radiography [1–3, 6] are reported. One set of problems involved the detection of lung lesions and mediastinal adenopathy from 12 bit per pixel (bpp) original CT images. The pixel intensities coded were not those of the original images, but rather of the residuals when pixel blocks are predicted by a simple Wiener-Hopf technique from previously encoded blocks. (See [5] for improvements that involve segmentation, increasing the memory of the predictor, and ridge regression.) Thirty images of each type were compressed to six different levels. Three radiologists then used the original and compressed images for diagnosis. We quantify outcomes by *sensitivity* (the chance an object is detected given that it is there) and *predictive value positive* (the chance that a detected object is actually there). Presumably larger bit rates are better, though the data do not bear this out for bit rates more than 2 bpp. Plots of outcome versus bit rate are fit by quadratic splines with a single knot and surrounded by bootstrap-based simultaneous confidence regions. On the basis of various analyses of the data we conclude that images can be compressed to bit rates between one and two bits per pixel without significant loss of diagnostic accuracy. From a different clinical study we have concluded that MR chest scans originally 9 bpp and used for measuring vessels in the chest can be compressed to .55 bpp without apparent loss of clinical accuracy [6]. Radiologists seem to like somewhat compressed images better than they like the originals [3].

References

[1] P.C. Cosman, C. Tseng, R.M. Gray, R.A. Olshen, L.E. Moses, H.C. Davidson, C.J. Bergin, and E.A. Riskin. "Tree-structured vector quantization of CT chest scans: Image quality and diagnostic accuracy," *IEEE Transactions on Medical Imaging*, vol. 12, pp. 727–739, December 1993.

[2] P.C. Cosman, H.C. Davidson, C.J. Bergin, C. Tseng, L.E. Moses, E.A. Riskin, R.A. Olshen, and R.M. Gray. "Thoracic CT images: Effect of lossy image compression on diagnostic accuracy of thoracic CT images," *Radiology*, vol. 190, pp. 517–524, February 1994.

[3] P.C. Cosman, R.M. Gray, and R.A. Olshen. "Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy," *Proceedings of the IEEE*, vol. 82, pp. 919–932, June 1994.

[4] A.B. Nobel and R.A. Olshen. "Termination and continuity of greedy growing for tree structured vector quantizers," Technical Report No. 164, Division of Biostatistics, Stanford University, Stanford, CA, December 1993. A revised version has been submitted for publication.

[5] G. Poggi and R.A. Olshen. "Pruned tree-structured vector quantization of medical images with segmentation and improved prediction," Technical Report No. 165, Division of Biostatistics, Stanford University, Stanford, CA, December 1993. To appear in revised form in *IEEE Transactions on Image Processing*, June 1995.

[6] C. Tseng, S.M. Perlmutter, P.C. Cosman, K.C.P. Li, C.J. Bergin, R.A. Olshen, and R.M. Gray. "Effect of tree-structured vector quantization on the accuracy of vessel measurements in MR chest scans," submitted for publication.

Greedy Growing of Tree-stuctured Classification Rules Using a Composite Splitting Criterion

Andrew B. Nobel[1]

Abstract — **We establish the Bayes risk consistency of an unsupervised greedy-growing algorithm that produces tree-structured classifiers from labeled training vectors. The algorithm employs a composite splitting criterion equal to a weighted sum of Bayes risk and Euclidean distortion.**

I. Introduction

Binary trees play an important role in the methodology of Statistics and Information Theory. Classification trees are used in a wide variety of statistical problems; tree-structured vector quantizers provide an efficient and effective means of compressing images.

A critical problem in practice is how to design a good tree-structured classifier or quantizer from a finite data set. Greedy growing algorithms [1,2,3] produce suitable trees one node at a time, optimizing a specified splitting criterion at each step. In spite of their empirical success, there has been little theory to support the unsupervised use of greedy growing algorithms, or to examine the behavior of such algorithms on large training sets.

We establish the Bayes risk consistency of an unsupervised variant of the CART [2] algorithm. The algorithm, which employs a composite splitting criterion equal to a weighted sum of Bayes risk and Euclidean distortion, is motivated by recent work [1] on the design of joint quantization/classification schemes. Variance of the classifiers is controlled by limiting the number of splits, rather than by pruning an 'overgrown' tree.

II. Definitions

A *tree-structured partition* is described by a pair (T, α) where T is a binary tree and $\alpha : T \to \mathbb{R}^d$ assigns a splitting vector to every node of T. Let $\tilde{T}$ denote the terminal nodes of T. Each vector $x \in \mathbb{R}^d$ is associated with a member of $\tilde{T}$ through a sequence of binary comparisons that trace a path through T: beginning at the root, and at each subsequent internal node, x moves to that child of the current node whose label is nearest to x in Euclidean distance. (A tie-beaking scheme may be used to avoid ambiguities.) Let V_t be the set of vectors x whose path contains the node t. Then each V_t is a convex polytope, and the the collection $\{V_t : t \in \tilde{T}\}$ is a partition of $\mathbb{R}^d$.

Let (X, Y) be jointly distributed random variables with $X \in \mathbb{R}^d$, $Y \in \{0, 1\}$. A *tree-structured classifier-quantizer* (TSCQ) is described by a four-tuple $(T, \alpha, \beta, \gamma)$, where (T, α) is a tree-structured partition as above, $\beta : T \to \mathbb{R}^d$ assigns a vector representative to each $t \in T$, and $\gamma : T \to \{0, 1\}$ assigns a class representative to each $t \in T$. (The four-tuple above will be abbreviated by T.) The triple (T, α, β) defines a tree-stuctured vector quantizer $Q_T = \sum_{t \in \tilde{T}} \beta(t) I\{x \in V_t\}$ by assigning a vector representative to each element of the partition $\{V_t : t \in \tilde{T}\}$. Similarly (T, α, γ) defines a tree-structured classification rule $C_T = \sum_{t \in \tilde{T}} \gamma(t) I\{x \in V_t\}$. Let $D(T) \triangleq E\|X - Q_T(X)\|^2$ be the distortion of Q_T and $R(T) \triangleq \mathbb{P}\{C_T(X) \neq Y\}$ the Bayes risk of C_T. Following [1], for $\lambda \in [0, 1]$, we define the *composite risk* $\Gamma_\lambda(T) = \lambda R(T) + (1 - \lambda) D(T)$.

III. Greedy Growing

Fix a TSCQ T and let $t \in \tilde{T}$. For each hyperplane L that intersects V_t° we may define an augmented tree $\hat{T} = \hat{T}(t, L)$ as follows: add children t_1, t_2 to t and select $\alpha(t_1), \alpha(t_2) \in V_t$ such that L is their perpendicular bisector; for $i = 1, 2$ let $\beta(t_i)$ be the Euclidean centroid of V_{t_i} and let $\gamma(t_i) = \text{argmin}_\theta \mathbb{P}\{Y = \theta | X \in V_{t_i}\}$. In this way $D(\hat{T}) \leq D(T)$ and $R(\hat{T}) \leq R(T)$, so that $\Gamma_\lambda(\hat{T}) \leq \Gamma_\lambda(T)$.

A *training sequence* $S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ consists of n independent replicas of (X, Y). Given S_n and an iteration count k_n the greedy growing algorithm produces a nested sequence $T_0 \leq T_1 \leq \ldots \leq T_{k_n}$ of TSCQ's. The initial tree T_0 consists of a single root node t_0 with $\alpha(t_o)$ arbitrary, $\beta(t_o) = 1/n \sum X_i$, and $\gamma(t_o)$ the majority vote among $\{Y_i : 1 \leq i \leq n\}$. Given T_r, the algorithm selects a terminal node $t^* \in \tilde{T}$ and a hyperplane L^* to minimize $\Gamma_\lambda(\hat{T}_r(t, L))$, and then sets $T_{r+1} = \hat{T}_r(t^*, L^*)$. All quantities are computed with respect to the empirical distribution of S_n. The output of the algorithm is T_{k_n}.

IV. Results

Let $R^* = \inf \mathbb{P}\{C(X) \neq Y\}$, where the infimum is taken over all classification rules $C : \mathbb{R}^d \to \{0, 1\}$. Unpruned classification trees produced by greedy growing are Bayes risk consistent if Euclidean distortion is given a non-zero weighting in the composite risk.

Theorem 1 *Let $\lambda < 1$ and suppose that the marginal distribution of X has a density such that $E\|X\|^2 < \infty$. For each $n \geq 1$ let T_n be produced by applying the greedy growing algorithm to the training sequence S_n for k_n steps. If (i) $k_n \to \infty$ and (ii) $n^{-1} k_n \log n \to 0$ then $R(T_n) \to R^*$ with probability one as $n \to \infty$.*

References

[1] R.M. Gray, KL. Oehler, K.O. Perlmutter, and R.A. Olshen, "Combining tree-structured vector quantization with classification and regression trees," *Proceedings of the 27th Asilomar Conference on Circuits, Systems, and Computers*, 1993.

[2] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Wadsworth International, Belmont, CA., 1984.

[3] P.A. Chou, *Applications of Information Theory to Pattern Recognition and the Design of Decision Trees and Trellises*, Ph.D. dissertation, Stanford University, 1988.

[1] Andrew Nobel is with the Department of Statistics, University of North Carolina, Chapel Hill. He is currently on leave at the Beckman Institute, University of Illinois, 405 N. Mathews, Urbana, IL 61801.

Tree-Structured Clustered Probability Models for Texture

Rosalind W. Picard and Kris Popat[1]
The Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA

Abstract — **A cluster-based probability model has been found to perform extremely well at capturing the complex structures in natural textures (e.g., better than Markov random field models). Its success is mainly due to its ability to handle high dimensionality, via large conditioning neighborhoods over multiple scales, and to generalize salient characteristics from limited training data. Imposing a tree structure on this model provides not only the benefit of reducing computational complexity, but also a new benefit – the trees are mutable, allowing us to mix and match models for different sources. This flexibility is of increasing importance in emerging applications such as database retrieval for sound, image and video.**

[1]This work was supported in part by NEC Corp. and HP Labs. We thank Tom Minka for discussions on merging models using trees.

Nonparametric Classifier Design Using Vector Quantization

Qiaobing Xie, Rabab K. Ward and Charles A. Laszlo
Dept. of Electrical Eng., Univ. of British Columbia, Vancouver, B. C., Canada V6T 1Z4

Abstract — **VQ-based method is developed as an effective data reduction technique for nonparametric classifier design. This new technique, while insisting on competitive classification accuracy, is found to overcome the usual disadvantage of traditional nonparametric classifiers of being computationally complex and of requiring large amounts of computer storage.**

I. Introduction

A solution to the excessive complexity problem of traditional nonparametric classifiers is to reduce the size of design set while insisting that the classifiers built upon the reduced design set should perform as well, or nearly as well as the classifiers built upon the original design set. This idea has been explicitly explored in the development of many classifier design algorithms using reduced sample sets. However, for very large design sets, these methods are often tedious and difficult to implement, and the final reduction rate is usually low [1].

We introduce a new approach for nonparametric data reduction using the vector quantization technique. Combining vector quantization with the classical Parzen's kernel and the *k*NN approaches, we develop two new algorithms of reduced nonparametric classifier design, which we shall denote the VQ-kernel and the VQ-*k*NN methods.

II. Development of VQ-based Nonparametric Classifiers

The philosophy guiding the development of most traditional nonparametric classification methods is that of using the statistical information contained in a set of pre-classified samples (or design set), for finding a good approximation of the actual underlying probability density function, $p(\mathbf{x})$. Then the classifier is built by applying the Bayesian rule. However, for achieving high classification performance, *this approximation to $p(\mathbf{x})$, while it is obviously sufficient, is not necessary*. For example, any good approximation to $[p(\mathbf{x})]^{\alpha}$, where constant $\alpha > 0$, will achieve the same Bayesian classifier as that achieved by approximating $p(\mathbf{x})$ itself.

In [2], Gersho shows that *for an optimal quantizer, in the asymptotic situation where the level of quantizer is sufficiently large, the density function of the reproduction vector will closely approximate a continuous density function $\lambda(\mathbf{x})$ which is proportional to $[p(\mathbf{x})]^{\alpha}$*, where α is a constant determined only by the dimension and the distance measure. This, along with our argument at the beginning of this section, strongly indicates that the reproduction alphabet in an optimal quantizer could be used as an effective design set for building classifiers.

The VQ-kernel Classifier: We propose that vector quantization be first applied to the original design set of each class. The reproduction alphabets of the resultant optimal quantizers are then retained as the reduced design sets. Then the kernel method is applied as usual except that the reduced design sets are used.

The VQ-kNN Classifier: Similar to the above VQ-kernel classifier except a *k*NN classifier is built with the reduced design sets.

IV. Simulations and Conclusions

By simulating with various data distributions, the performance of our VQ-based methods are compared with that of the traditional reduction algorithms including the CNN, RNN, ENN, ECNN, as well as Fukunaga's reduced Parzen [1]. Fig. 1 shows error rates for real speech data.

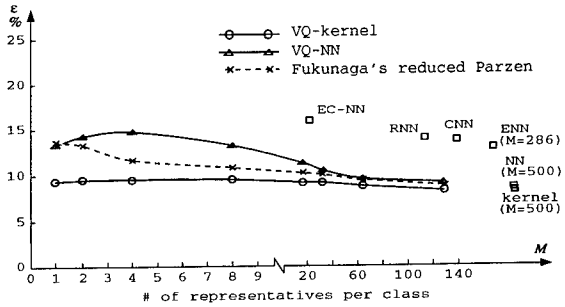


Fig. 1. Classification error rates of VQ-based classifiers and traditional classifiers for speech data.

It is found that 1): the VQ-kernel classifier outperforms, in terms of accuracy, all other data reduced algorithms at all the reduction levels; 2) the VQ-based methods usually achieve tens of times higher reduction rates while giving the same level of classification accuracy — this usually means a drastic reduction in complexity and storage; 3) finding the reduced design set in the VQ-based methods is tens even hundreds times faster than that in other proposed data reduction algorithms [1].

The VQ-based classifier design technique is also extended to the design of histogram-based classifiers [3].

References

[1] Q. Xie, C. A. Laszlo, and R. K. Ward, "Vector quantization technique for nonparametric classifier design," *IEEE Trans. on Pattern Anal. Machine Intel.*, vol. 15, pp. 1326–30, Dec. 1993.

[2] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. on Inform. Theory*, vol. IT-25, pp. 373–80, July 1979.

[3] Q. Xie, R. K. Ward, and C. A. Laszlo, "Multidimensional histogram classifier design by using vector quantization," in *Proceedings of IEEE Pacific Rim Conf. on Comm., Comp. Signal Processing*, vol. 1, (Vancouver), pp. 39–42, Sept. 1993.

Tree-Based Models for Speech and Language

Michael D. Riley

AT&T Bell Laboratories, 600 Mountain Ave., Murray Hill, NJ 07974

Abstract – **Several applications of statistical tree-based modelling are described to problems in speech and language, including prediction of possible phonetic realizations, segment duration modelling in speech synthesis and end of sentence detection in text analysis.**

I. INTRODUCTION

Classification and regression trees [1] are well suited to many of the pattern recognition problems encountered in speech and language since they (1) statistically select the most significant features involved (2) permit both categorical and continuous factors to be considered, (3) provide "honest" estimates of their performance, and (4) allow human interpretation and exploration of their result. Below we describe several applications of these methods to speech and language processing.

II. PREDICTION OF POSSIBLE PHONETIC REALIZATIONS

A lattice of possible close phonetic transcriptions given a phonemic transcription (from the orthography and a dictionary) is produced using a 6000 sentence, multispeaker transcribed database as input. The resulting phonetic network predicts the correct pronunciation of a phoneme on test data from the same corpus 83% of the time, contains the correct phone in the top 5 guesses 99% of the time, and has a conditional entropy of .8 bits. This compares to the null model, in which only the phoneme to realize is used, that predicts the correct phone 69% of the time, contains the correct phone in the top 10 guesses 99% of the time, and has a conditional entropy of 1.5 bits. [2]

III. SEGMENT DURATION MODELLING IN SPEECH SYNTHESIS

400 utterances from a single speaker and 4000 utterances from 400 speakers of American English were used to build optimal decision trees that predict segment durations. Over 70% of the durational variance for the single speaker and over 60% for the multiple speakers were accounted for by this method when using information only at the word level and below. These trees were used to derive durations for a text-to-speech synthesizer and were found to give results comparable to the existing heuristically derived duration rules. Since tree building and evaluation is rapid once the data are collected and the candidate features specified, the technique can be readily applied to other feature sets and to other languages. [3]

IV. END OF SENTENCE DETECTION

The not-so-simple problem of deciding when a period in text corresponds to the end of a declarative sentence (and not an abbrev.) is attempted with trees using the Brown corpus as input. The result is 99.8% correct classification. The many special cases required to solve this problem well, nicely show the value of the tree approach here. The majority of the errors are due to difficult cases, e.g. a sentence that ends with "Mrs." or begins with a numeral [4].

V. DISCUSSION

On the whole, we have found classification and regression trees quite useful in modelling a variety of phenonema in speech and language. In part, it is their ability to handle both categorical and continuous inputs and outputs that makes them attractive to us. The fact that they offer efficient algorithms, a well-established cross-validation procedure, and a relatively perspicuous representation makes them more appealing to us than, say, back-propogation neural networks for the problems we have described.

The principal difficulty we have found with this and similar statistical approaches is that while the trees classify well most of the time, they occasionally make egregious errors. When noticed, it is possible to correct these errors by hand modification of the trees. This is, however, quite tedious. Further, if new data are used or new input features are tried, the editing has to be redone (if the error remains).

What would be most appealing to us would be techniques that would allow easy mixing of statistical learning with hand specification. The user could hand specify what he is sure of and leave to the statistics to fill in the rest the best it can, letting us have our cake and eat it too.

REFERENCES

[1] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks, Pacific Grove CA, 1984.

[2] Michael Riley. A statistical model for generating pronunciation networks. In *Proceedings of the Speech and Natural Language Workshop*, page S11.1. DARPA, Morgan Kaufmann, October 1991.

[3] Michael Riley. Tree-based modelling of segmental durations. In G. Bailly and C Benoit, editors, *Talking Machines*, pages 265–274. North-Holland, 1992.

[4] Michael Riley. Some applications of tree-based modelling to speech and language. In *Proceedings of the Speech and Natural Language Workshop*, pages 339–352, Cape Cod MA, October 1989. DARPA, Morgan Kaufmann.

Image Coding via Bintree Segmentation and Texture VQ

Xiaolin Wu[1]

Dept. of Computer Science, Univ. of Western Ontario, London, Ontario, CANADA, N6A 5B7

Image compression is often approached from an angle of statistical image classification. For instance, VQ-based image coding methods compress image data by classifying image blocks into representative two-dimensional patterns (codewords) that statistically approximate the original data. Another image compression approach that naturally relates to image classification is segmentation-based image coding (SIC). In SIC, we classify pixels into segments of certain uniformity or similarity, and then encode the segmentation geometry and the attributes of the segments.

Image segmentation in SIC has to meet some more stringent requirements than in other applications such as computer vision and pattern recognition. Firstly, the segmentation description must be compact to ensure low bit rate. Secondly, the classification criterion should quantify visual differentiation of image patterns. Thirdly, the segmentation process has to be fast enough to suit image/video coding purposes.

An efficient SIC coder has to strike a good balance between accurate semantics and succinct syntax of the segmentation. From a pure classification point of view, free form segmentation by relaxation, region-growing, or split-and-merge techniques offers an accurate boundary representation. But the resulting segmentation geometry is often too complex to have a compact description, defeating the purpose of image compression. Instead, we adopt a bintree-structured segmentation scheme. The bintree is a binary tree created by recursive rectilinear bipartition of an image. The bintree-structured segmentation is semantically more flexible than the popular quadtree, and yet it has as simple syntax as the quadtree. This nice property translates to compression gains.

Large and smooth surfaces of a natural scene correspond to regions of fairly continuous intensities in its digital image. In these regions pixel values can be fit well by a low-order polynomial, and the least-square piecewise functional approximation yields more compact image description than DCT and VQ. Luckily, the majority areas of a natural image fall into this category. This is demonstrated by the facts that most codewords in a VQ codebook form smooth shading patterns, and most DCT blocks have dominant low frequency coefficients. The main advantage of SIC over VQ and DCT is in that it can adaptively fit an image with more flexible segments than fixed blocks, resulting in fewer segments hence shorter description. However, in the areas of textures or edges, least-square piecewise fitting breaks down since higher order terms induce more real coefficients to be quantized and coded. In comparison, VQ technique is more suitable to classify and compress textures. Thus within a SIC framework, least-square approximation and VQ can complement each other for higher compression than either method alone. This necessitates the classification of an image into smooth and texture regions.

Many possible classifiers can be used to decide whether a bintree block is smooth or contains textures or edges. Since we use least-square approximation to code smooth regions, it is convenient to base the classifier on variance. We fit pixel values by a low-order polynomial to form a largest bintree block possible under a given error tolerance. The size of the bintree block serves as the classifier. If the size exceeds a threshold, we hypothesize that the intensity function is smooth in that area, and consequently code the block with quantized polynomial coefficients. Otherwise, we hypothesize that the block contains rich textures. An additional VQ texture coding is employed on the block to get a better approximation. Using bintree block size as the classifier means that no side information is required to identify the type of the segment.

To design a segmentation algorithm, we can either split the image top-down or merge primitive blocks bottom-up. But there are two advantages to the bottom-up merge approach. By merging smaller blocks, we do not unnecessarily solve the least-squares problem in large bintree blocks which cannot possibly be leaf nodes, reducing algorithm complexity. Also, by examining smaller blocks first, we avoid segment misclassification due to smoothing of prominent local textures by least-square fitting.

A main result of this research is a texture code based on binary VQ. Let $f(x,y)$ be the input image and $g(x,y)$ be the least-square linear approximation of $f(x,y)$ in a texture block. We model the texture to be $e(x,y) = f(x,y) - g(x,y)$. DCT or VQ can be used to encode $e(x,y)$. But lower bit rate can be achieved for the same transparent image quality by taking the advantage of the fact that in high texture areas human visual system is less sensitive to intensity resolution. Therefore, we coarsely quantize the amplitude of $e(x,y)$ into only two levels, and map $e(x,y)$ to a binary texture pattern $m(x,y)$, where $m(x,y) = 0$ if $e(x,y) \geq 0$ and $m(x,y) = 1$ if $e(x,y) < 0$. Depending on $m(x,y) = 0$ or $m(x,y) = 1$, $e(x,y)$ is quantized to $-\sigma$ or σ, where σ is the standard deviation of $e(x,y)$. Note that $e(x,y)$ is zero mean since it is the residual function of a least-square linear approximation. Consequently, this simple bi-level quantization preserves both mean and variance of the original image just like in block truncation coding.

To obtain rates lower than 1 bit/pixel, we have to compress the texture pattern $m(x,y)$ as well. This is a problem of texture classification which can be solved by binary VQ. But a direct use of the LBG algorithm often fails to produce a satisfactory codebook, because the number of local optima in the sample space of binary vectors is too numerous for a gradient-descent optimization method to improve on an initial codebook. To worsen the problem, the expected Hamming distance previously used in binary VQ as the distortion measure does not proportionally quantify the visual quality degradation caused by inverted bits. This is because the perceived texture reproduction quality is affected not only by the average error but also by burst errors. It is important not to invert two or more adjacent bits. We discover that the use of linear codes for binary VQ in the spirit of optimal sphere covering offers remedies to both problems of local minimum trap and burst error.

[1]This work was supported by the Natural Sciences and Engineering Research Council of Canada.

SESSION III

Randomization Complexity and Information Theory

Coding for Noisy Feasible Channels

Richard J. Lipton[†]

Department of Computer Science

Princeton University

Princeton, NJ 08544

rjl@princeton.edu

Abstract: We prove a constructive version of Shannon's Fundamental Theorem of Information Theory. The new theorem holds for any *feasible* channel. A channel is feasible provided it is computable by a polynomial time computation.

Our main result is a new constructive proof of Shannon's Theorem. Consider a feasible channel. Then, there is a coding method C with the following properties:

(*1*) We can construct C in polynomial time.

(*2*) We can encode any message in polynomial time.

(*3*) We can decode any message in polynomial time.

(*4*) The probability that the method makes an error goes to 0 at least as fast as $1/n^{O(1)}$.

(*5*) The rate of the method can be as close to the capacity of the channel as one wishes.

How do we construct these codes? Following Lipton [1] we restrict the channel to be "feasible". That is we restrict the channel to only use random polynomial time to decide which bits to change. Essentially, any channel is characterized by two parameters: (i) how "mean" it is; (ii) how "smart" it is. We measure how mean a channel is by how *many* bits it can change. We measure how smart a channel is by how much computation it is allowed to perform to decide *which* bits to change. Thus, our key point is: only allow channels with smartness bounded by random polynomial time.

We claim that "real" channels have their smartness limited in this way. This is an assertion like "Church's Thesis" and of course cannot be "proved". It is, we claim, a reasonable assumption. Real channels are analog/digital systems. It certainly appears to be reasonable to assume that such systems cannot do more computing than random polynomial time. If a real channel existed that could do more than polynomial time computation, then perhaps we could use it to solve intractable problems! Note, in classic information theory often the most powerful kind of channel considered is a channel that is finite state. Of course this is in our class.

References

[1] R. Lipton, *A New Approach to Information Theory*, in STACS 1994.

[†] Supported in part by NSF CCR-9304718

Coding for Distributed Computation

Leonard J. Schulman*
Computer Science Division
U. C. Berkeley
Berkeley CA 94720

Extended communications among component processors are essential to the operation of all but the simplest computers. In this talk we are concerned with the following question: if the communications among processors, linked in some network, are unreliable, what is the effect on the efficiency and reliability with which the network can perform a computation?

An important case of this scenario, that in which there are two processors and the required task is to transmit a large block of data from one to the other, actually predates large-scale computing. Shannon's coding theorem addresses this problem, and shows that in order to reliably transmit a message of T bits over a noisy communication channel it suffices to send a message of length $T\frac{1}{C}$ (for $0 < C < 1$ the "Shannon capacity" of the channel). The theorem ensures that the probability of a decoding error is exponentially small in the message length T.

We will describe analogous coding theorems for the more general, interactive, communications required in computation. In this case the bits transmitted in the protocol are not known to the processors in advance but are determined dynamically. Therefore the block encoding technique used in the proof of Shannon's theorem, does not apply.

First we show that any interactive protocol of length T between two processors connected by a noiseless channel can be simulated, if the channel is noisy (a binary symmetric channel of capacity C), in time proportional to $T\frac{1}{C}$, and with error probability exponentially small in T.

Then we show that this result can be extended to arbitrary distributed network protocols. We show that any distributed protocol which runs in time T on a network of degree d having noiseless communication channels, can, if the channels are in fact noisy, be simulated on that network in time proportional to $T\frac{1}{C}\log d$. The probability of failure of the protocol is exponentially small in T.

Preliminary presentations of these results can be found in [1, 2].

The network theorem is joint with Sridhar Rajagopalan.

*Research supported by an NSF Mathematical Sciences Postdoctoral Fellowship.

References

[1] S. Rajagopalan and L. J. Schulman. Coding for distributed computation. In *Proceedings of the 26th Annual Symposium on Theory of Computing*, 1994.

[2] L. J. Schulman. Deterministic coding for interactive communication. In *Proceedings of the 25th Annual Symposium on Theory of Computing*, pages 747–756, 1993.

Minimal Randomness and Information Theory

Sergio Verdú
Dept. Electrical Engineering, Princeton University, Princeton, New Jersey 08544, USA

Abstract — **This is a tutorial survey of recent information theoretic results dealing with the minimal randomness necessary for the generation of random processes with prescribed distributions.**

I. Introduction

Shannon Theory explores the fundamental limits on the size of codes that enable the reliable reproduction or transmission of information. Reliability is typically quantified by the probability that the decoded message is equal to the original one, or by some measure of the distance between the original and decoded messages.

In this paper we are not interested in the reproduction or transmission of information but rather in the generation of random processes with prescribed distributions, and associated problems. For example, we may want to simulate a "real-world" random process, or the response of a system to such an input. Random process generation is accomplished by adequately mapping a source of pure random bits. A key question that quantifies the "complexity" of the random process is the *minimal randomness* of the source of pure bits necessary to accomplish the task. As in conventional Shannon theory, a rich theory arises when some distance (often arbitrarily small) is allowed between the desired and the resulting probability distributions. To this end, several distance measures have been considered in the literature, such as variational distance, divergence, ρ-distance, etc.

II. Source Resolvability

The resolvability of a source is defined [1] as the minimal number of random bits per sample it takes to reproduce the n-dimensional distributions with arbitrary accuracy as n tends to infinity. A general formula is shown in [1] for the resolvability of a source. For the special case of stationary ergodic sources and variational distance it is equal to the entropy rate. Reference [7] considers the problem of finite-precision resolvability where the approximation distance need not be arbitrarily small, and shows that for any information stable source D-resolvability is independent of D in the special case of variational distance. However, with less stringent approximation measures such as the Prohorov and ρ-distance, the D-resolvability is shown in [7] to be given by the rate distortion function evaluated with a sample-path distortion metric derived from the distribution distance measure.

III. Channel Resolvability

In system simulation, the objective is to induce the same output distributions as those that would obtain with a "real-world" input. The channel (or system) resolvability defined as the minimal randomness required to generate any desired input so that the output distributions are approximated with arbitrary accuracy. Naturally, the more "random" a system is, the lower its resolvability, as it does not pay to reproduce fine details in the input distributions. It is shown in [1] that the channel resolvability is equal to its capacity for most discrete channels (those that satisfy the strong converse). The complementary problem where the input is given but the channel is to be simulated is studied in [5], where it is shown that the minimal randomness required to simulate the system for a specific input is equal to the conditional entropy rate of the output given the input.

IV. Intrinsic Randomness

A problem which is dual to source resolvability is the *maximal* randomness rate that can be extracted from an arbitrary source. The *intrinsic randomness rate* of a source is defined in [9] as the largest rate of *almost-fair* coin flips that can be extracted by a deterministic mapping of the source. For stationary ergodic sources and variational distance the intrinsic randomness rate is equal to the entropy rate [9]. However there are nonstationary sources for which the intrinsic randomness rate is not equal to the minimal noiseless source coding rate. The more general problem of finite precision intrinsic randomness is studied in [10]. Using variational distance, [10] shows that the finite precision intrinsic randomness rate is given, essentially, by the inverse asymptotic distribution of the entropy density.

References

[1] T. S. Han, S. Verdú, "Approximation Theory of Output Statistics," *IEEE Trans. on Information Theory,* vol. IT-39, pp. 752-772, May 1993.

[2] T. S. Han, S. Verdú, "Spectrum Invariancy under Output Approximation for Discrete Memoryless Channels with Full Rank," *Problemii Peredachi Informatsii,* (in Russian), vol. 29, no. 2, p. 9-27, 1993, translated in *Problems of Information Transmission,* Apr.-June 1993, p. 101-118.

[3] T. S. Han and S. Verdú, "The Resolvability and the Capacity of the AWGN Channel are equal," *Proc. 1994 IEEE Int. Symp. Information Theory,* Trondheim, Norway, June 1994, p. 463.

[4] M. Burnashev and S. Verdú, "Measures separated in L_1 Metrics and ID-codes," *Problemii Peredachi Informatsii,* (in Russian), to appear.

[5] Y. Steinberg, S. Verdú, "Channel Simulation and Coding with Side Information," *IEEE Trans. on Information Theory,* vol. 40, no. 3, pp. 634-646, May 1994.

[6] Y. Steinberg and S. Verdú, "The Random Bit Rate required for Channel Simulation," *Proc. Sixth Joint Swedish-Russian International Workshop on Information Theory,* pp. 447-451, Moelle, Sweden, Aug. 22-27, 1993.

[7] Y. Steinberg and S. Verdú, "Coarse Approximations of Source Statistics and Rate-Distortion Theory," *IEEE Trans. on Information Theory,* submitted.

[8] Y. Steinberg and S. Verdú, "Finite Precision Source Resolvability," *Proc. 1994 IEEE Int. Symp. Information Theory,* Trondheim, Norway, June 1994, p. 296.

[9] S. Vembu and S. Verdú, "Generating Random Bits from an Arbitrary Source: Fundamental Limits," *IEEE Trans. on Information Theory,* submitted.

[10] Y. Steinberg and S. Verdú, "Finite Precision Intrinsic Randomness and Source Resolvability," *Proc. IT/STAT Workshop '94,* Alexandria, VA, Oct. 1994.

Finite-precision Intrinsic Randomness and Source Resolvability

Yossef Steinberg and Sergio Verdú

C3I Center, George Mason Univ., Fairfax, VA 22030 and Dept. of EE, Princeton Univ., Princeton, NJ 08544, USA

I. Introduction and Definitions

Random number generators are important devices in randomized algorithms, Monte-Carlo methods, and in simulation studies of random systems. A random number generator is usually modeled as a random source emitting independent, equally likely random bits. In practice, the random source one has at hand can deviate from this idealized model, and the random number generator operates by applying a deterministic mapping on the output of the (nonideal) random source. The deterministic mapping is chosen so that the resulting process approximates – in some sense – a sequence of independent, equally likely random bits. A prime measure of the intrinsic randomness of a given source X is the maximal rate at which random bits can be extracted from X by suitably mapping its output. This maximal rate depends on the statistics of the source X and on the sense of approximation. In [1] it is shown that the maximal rate at which arbitrarily accurate approximations of pure random bits can be extracted from X equals its inf entropy rate, $\underline{H}(X)$. The measures of accuracy with respect to which this result was shown to hold are the variational distance, the $\bar{d}$ distance and normalized divergence.

In problems like randomized algorithms, or Monte-Carlo simulations, an arbitrarily accurate approximation of pure random bits may be more than what we need, and a controlled deviation from pure random bits can be tolerated. In such cases, one may wish to increase the rate of generation of random bits at the expense of a coarser approximation of the desired fair coin flip distributions. In this work we study the problem of finite-precision random bit generation, where the accuracy measure is the variational distance. The results presented here extend part of the results in [1] and also provide a nice counterpart to the finite-precision source resolvability problem that was studied in detail in [2].

Throughout, X is a random source with finite alphabet A, and logarithms have base 2. We start with a few definitions.

Definition 1 [1] R is a *D-achievable intrinsic randomness rate of X* if there exists a sequence of deterministic mappings $\phi_n : A^n \to \{0,1\}^r$ such that for all $\gamma > 0$ and sufficiently large n,

$$\frac{r}{n} > R - \gamma$$

and

$$d_v(\phi_n(X^n), B^r) \le D$$

where B^r stands for an equiprobable distribution over $\{0,1\}^r$ and $d_v(\cdot,\cdot)$ is the variational distance between distributions.

Definition 2 The *finite-precision intrinsic randomness rate of X* is defined as the supremum of the D-achievable intrinsic randomness rates of X and is denoted by $U_v(D,X)$.

Note that $U_v(2,X) = \infty$ for every source X. The next definition deals with the relevant information theoretic function.

Definition 3 The *variational inf rate-distortion function of X*, $\underline{R}_v(D)$, is defined as the supremum over all real numbers h satisfying

$$\limsup_{n\to\infty} P_X^n\left(\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} < h\right) \le \frac{D}{2}.$$

Thus, $\underline{R}_v(D)$ is the largest real number h such that the mass of the entropy density to the left of h does not exceed $D/2$, asymptotically. Note that for every source $\underline{R}_v(0)$ equals the inf entropy-rate of the source, $\underline{H}(X)$, and $\underline{R}_v(2) = \infty$.

II. Results

Theorem 1

$$U_v(D,X) = \underline{R}_v(D).$$

The next corollary is an easy consequence of Definition 3 and Theorem 1: it implies that if X is information stable, one cannot increase the asymptotic rate of production of random bits by increasing their deviation (w.r.t. variational distance) from ideal fair coin flips. This result has a nice counterpart in the finite-precision source resolvability problem: it is shown in [2] that if X is information stable, then its variational finite-precision resolvability $S_v(D,X)$ is independent of D in the region $0 < D < 2$.

Corollary 1 *If X is information stable, then for $0 < D < 2$*

$$\underline{R}_v(D) = U_v(D,X) = H(X).$$

In [2] the variational finite-precision source resolvability was characterized as the infimum of the sup information rate over an appropriate class of channels — the corresponding sup rate-distortion function. The nice duality between the problems of finite-precision source resolvability and finite-precision bit generation, and Corollary 1, leads one to suspect that the variational finite-precision source resolvability (and hence also the sup rate-distortion function) as defined in [2] admits a simpler characterization - such as that in Definition 3. This is indeed the case, as one can see from the following theorem.

Theorem 2

$$\begin{aligned} S_v(D,X) &= \overline{R}_v(D) \\ &= \inf\left\{h : \limsup_{n\to\infty} P_{X^n}\left(\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} > h\right) \le \frac{D}{2}\right\}. \end{aligned}$$

Thus, the variational sup rate-distortion function as defined in [2] is also equal to the smallest real number h such that the mass of the entropy density to the *right* of h does not exceed $D/2$, asymptotically.

References

[1] S. Vembu, S. Verdú, "Generating Random Bits from an Arbitrary Source: Fundamental Limits," submitted.

[2] Y. Steinberg, S. Verdú, "Coarse Approximations of Source Statistics and Rate-Distortion Theory," submitted.

Identification via Compressed Data

Rudolf Ahlswede[1], En-hui Yang, and Zhen Zhang[2]

I. Introduction

In this paper, a combined problem of source coding and identification is considered. To put our problem in perspective, let us first review the traditional problem in source coding theory. Consider the following diagram, where $\{X_n\}_{n=1}^{\infty}$ is an

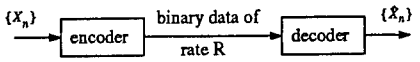


Figure 1: Model for source coding

i.i.d source taking values on a finite alphabet $\mathcal{X}$. The encoder output is a binary sequence which appears at a rate R bits per symbol. The decoder output is a sequence $\{\hat{X}_n\}_1^{\infty}$ which take values on a finite reproduction alphabet $\mathcal{Y}$. In traditional source coding theory, the decoder is required to be able to recover $\{X^n\}_1^{\infty}$ completely or with some allowable distortion. That is, the output $\{\hat{X}_n\}_1^{\infty}$ must satisfy

$$n^{-1}\sum_{i=1}^{n}\rho(X_i,\hat{X}_i)\le d \qquad (1)$$

for sufficiently large n, where $\rho:\mathcal{X}\times\mathcal{Y}\to[0,+\infty)$ is a distortion measure and $d\ge 0$ is the allowable distortion. The problem is then to determine the infimum of rate R such that the system shown in Fig.1 can operate in such a way that (1) is satisfied. From rate distortion theory, this infimum is given by the rate distortion function of the source $\{X_n\}_1^{\infty}$.

Let us now consider the system shown in Fig. 2. The se-

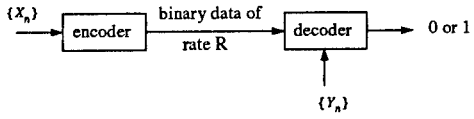


Figure 2: Model for joint source coding and identification.

quence $\{Y_n\}_1^{\infty}$ is a sequence of i.i.d random variables taking values on $\mathcal{Y}$. Known $\{Y_n\}$, the decoder is now required to be able to identify whether or not the distortion between $\{X_n\}$ and $\{Y_n\}$ is less than or equal to d in such a way that two kinds of error probabilities satisfy some prescribed conditions. The problem we are now interested in is still to determine the infimum of rate R such that the system shown in Fig.2 can operate in this way.

II. Formal Formulation of Problem

Let $\{(X_n,Y_n)\}_1^{\infty}$ be a sequence of independent drawings of a pair (X,Y) of random variables taking values on $\mathcal{X}\times\mathcal{Y}$ with joint distribution P_{XY}. Fix $0\le d<E\rho(X,Y)$. An nth-order identification (ID) code $\mathcal{C}_n$ is defined to be a triple $\mathcal{C}_n=(f_n,B_n,g_n)$, where $B_n\subset\{0,1\}^*$ is a prefix set, f_n(called an "encoder") is a mapping from $\mathcal{X}^n$ to B_n, and g_n(called a "decoder") is a mapping from $\mathcal{Y}^n\times B_n\to\{0,1\}$. When $\mathcal{C}_n$ is used in the system shown in Fig.2, its performance can be measured by the following three quantities: the resulting average rate defined by $r_n(\mathcal{C}_n)=En^{-1}$(the length of $f_n(X^n)$), the first kind of error probability defined by $p_{e1}(\mathcal{C}_n)=\Pr\{g_n(Y^n,f_n(X^n))=0|\rho_n(X^n,Y^n)\le d\}$, and the second of error probability defined by $p_{e2}=\Pr\{g_n(Y^n,f_n(X^n))=1|\rho_n(X^n,Y^n)>d\}$.

Let $R\in[0,+\infty)$, $\alpha\in(0,+\infty]$ and $\beta\in(0,+\infty]$. A triple (R,α,β) is said to be achievable if for any $\epsilon>0$, there exists a sequence $\{\mathcal{C}_n\}$ of ID codes, where $\mathcal{C}_n=(f_n,B_n,g_n)$ is an nth-order ID code, such that for sufficiently large n,

$$r_n(\mathcal{C}_n)\le R+\epsilon\,,\quad p_{e1}\le 2^{-n(\alpha-\epsilon)}\ \text{ and }\ p_{e2}\le 2^{-n(\beta-\epsilon)}\,,$$

where as a convention, $\alpha=+\infty(\beta=+\infty$, resp.) means that the first(second, resp.) kind of error probability of $\mathcal{C}_n$ is equal to 0. Let $\mathcal{R}$ denote the set of all achievable triples. In this paper, we are interested in determining the closure $\bar{\mathcal{R}}$ of $\mathcal{R}$. Specifically, we define for each pair (α,β), where $\alpha,\ \beta\in[0,+\infty]$,

$$R^*_{XY}(\alpha,\beta,d)=\inf\{R|(R,\alpha,\beta)\in\bar{\mathcal{R}}\}\,.$$

Our main problem is then the determination of the function $R^*_{XY}(\alpha,\beta,d)$.

III. Main Results

Assume that X and Y are independent. For any $0<d<E\rho(X,Y)$, define $\beta(d)$ by $\beta(d)=\inf D(P||P_{XY})$, where the infimum is taken over all distributions P on $\mathcal{X}\times\mathcal{Y}$ such that $\sum_{x,y}P(x,y)\rho(x,y)\le d$. Let U be a random variable taking values on some finite set $\mathcal{U}$. Let P_{XU} denote the joint distribution of X and U. For any $\alpha\ge 0$, define

$$\mathcal{E}(P_{XU},\alpha,d)=\inf\{D(P_{\tilde{Y}}||P_Y)+I(U;\tilde{Y})\}\,,$$

where the infimum is taken over all random variables $\tilde{Y}$ taking values on $\mathcal{Y}$ such that $E\rho(X,\tilde{Y})\le d$ and $D(P_{\tilde{Y}}||P_Y)+I(XU;\tilde{Y})\le\beta(d)+\alpha$. Here we make use of the convetion that the infimum taken over an empty set is $+\infty$. We define for any $\beta>0$

$$R(P_X,P_Y,\alpha,\beta,d)=\inf\{I(X;U)|U\text{is a R.V. with }\mathcal{E}(P_{XU},\alpha,d)\ge\beta\}$$

and let

$$R(P_X,P_Y,\alpha,0,d)=\lim_{\beta\to 0+}R(P_X,P_Y,\alpha,\beta,d)\,.$$

The following theorem gives a general formula for $R^*_{XY}(\alpha,\beta,d)$.

Theorem 1 *For any $0<d<E\rho(X,Y)$, $0\le\beta<\beta(d)$, and $\alpha\in(0,+\infty]$, the following holds*

$$R^*_{XY}(\alpha,\beta,d)=\bar{R}(P_X,P_Y,\alpha,\beta,d)\,,$$

where

$$\bar{R}(P_X,P_Y,\alpha,\beta,d)=\lim_{\beta'\to\beta-}R(P_X,P_Y,\alpha,\beta',d)\,.$$

The converse part of Theorem 1 is related to the general isoperimetric problem. During the process of proving the converse part, we develope a new powerful method for converse-proving in multi-user information theory. For more details, please refer to [1].

[1]Fakultaet fuer Mathematik, Universitaet Bielefeld, 4800 Bielefeld 1, Germany

[2]Commun. Science Institute, Dept. of EE-Systems, University of Southern California, Los Angeles, CA 90089-2565.

References

[1] R. Ahlswede, E.-H. Yang and Z. Zhang, "Identification via compressed data," Preprint, 1994.

Testing of Composite Hypotheses and ID–codes

M.V.Burnashev and S.Verdu

Abstract – A geometrical approach to ID–codes, based on their equivalence to some natural notions from mathematical statistics is described. That not only enlarges the available analytical apparatus, but also enables us to strengthen some known results.

Let A and B be finite input and output alphabets of a stationary memoryless channel with conditional transition probabilities $W(b|a), a \in A, b \in B$. If P is some probability distribution (measure) on the channel input A^n then by $Q = PW^{(n)}$ we denote the generated distribution on the channel output B^n.

Definition 1 [1]. A collection $(P_i, \mathcal{D}_i, i = 1, \ldots, M)$ of probability measures P_i on A^n and regions $\mathcal{D}_i \subseteq B^n$ is called an (M, n, δ) – ID–code if the following conditions are satisfied:

$$Q_i(\mathcal{D}_i) \geq 1 - \delta \text{ and } Q_i(\mathcal{D}_j) \leq \delta \text{ for any } i \neq j.$$

What concerns the maximal cardinality $M(n, \delta)$ of ID–codes, it is known that [1,2]

$$\lim_{n \to \infty} \frac{\ln \ln M(n, \delta)}{n} = C, \quad 0 < \delta \leq \delta_0, \qquad (1)$$

where C – channel capacity and δ_0 is some positive constant.

Another meaning of Definition 1 is that the collection of measures $\{P_i\}$ of an ID–code has the following property: any simple hypotheses P_i can be "tested" against the composite alternative consisting of all remaining measures $\{P_j,\ j \neq i\}$ from the same family. Or, any measure P_i is "almost orthogonal" to the convex combination of all remaining measures.

We develop [3] in a quantative manner this connection between ID–codes and the testing of composite hypotheses. Such an approach not only enlarges the research analytical apparatus, but also enables us to strengthen some results from [1,2]. In particular, it is shown that the equality (1) remains valid for any $0 \leq \delta < 1/2$. That gives certain completeness to (1), since for $\delta \geq 1/2$ the number $M(n, \delta)$ becomes infinite provided that randomized decision rules are allowed for use.

References

[1] R. Ahlswede and G. Dueck, "Identification via Channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 15–29, 1989.

[2] T. S. Han and S. Verdu, "New Results in the Theory of Identification via Channels," *IEEE Trans. Inform. Theory*, vol. 38, pp. 14–25, 1992.

[3] M. V. Burnashev and S. Verdu, "Measures separated in L–1 metrics and ID-codes," *Probl. Inform. Trans.*, vol. 30, No. 3, pp. 3–14, 1994.

[4] L. A. Bassalygo and M. V. Burnashev, "Estimate for the maximal number of messages for a given probability of succesful deception," *Probl. Inform. Trans.*, vol. 30, No. 2, pp. 42–48, 1994.

SESSION IV

Nonparametric Function Estimation

Asymptotically Optimal Model Selection and Neural Nets

Andrew R. Barron
Statistics Dept., Yale University, P.O. Box 208290, New Haven, CT 06520, USA

Abstract — **A minimum description length criterion for inference of functions in both parametric and nonparametric settings is determined. By adapting the parameter precision, a description length criterion can take on the form $-\log(\text{likelihood}) + \text{const} \cdot m$ instead of the familiar $-\log(\text{likelihood}) + (m/2)\log n$ where m is the number of parameters and n is the sample size. For certain regular models the criterion yields asymptotically optimal rates for coding redundancy and statistical risk. Moreover, the convergence is adaptive in the sense that the rates are simultaneously minimax optimal in various parametric and nonparametric function classes without prior knowledge of which function class contains the true function. This one criterion combines positive benefits of information-theoretic criteria proposed by Rissanen, Akaike, and Schwarz. It is also reviewed how the minimum description length principle provides accurate estimates in irregular models such as neural nets.**

I. Minimum description length criterion

Data $Y_1, Y_2, ..., Y_n$ are assumed to be independent with an unknown density p. Let a sequence of parametric families be given. Each family has a density $p_k(y|\theta)$, parameter space Θ_k, and codelengths $L(k)$, $L(\theta|k)$ for the model index k and parameters θ in a discrete subset $\tilde{\Theta}_k \subset \Theta_k$. The codelengths are assumed to satisfy Kraft's inequality. Then $\min_{\theta\in\tilde{\Theta}_k}\{\log 1/p_k(Y^n|\theta) + L(\theta|k) + L(k)\}$ is the length of a uniquely decodable code for the data, where $p_k(Y^n|\theta) = \prod_{i=1}^n p_k(Y_i|\theta)$. The index $\hat{k}$ and parameter value $\hat{\theta}$ achieving the minimum description length (MDL) provides the density estimator $\hat{p}(y) = p_{\hat{k}}(y|\hat{\theta})$ [2,3].

The data compression quality is measured by the redundancy of the MDL code, which is bounded by the index of resolvability $R_n(p) = \min_{k,\theta}\{D(p||p_{k,\theta}) + (1/n)(L(\theta|k) + L(k))\}$, where $D(p||q)$ denotes the Kullback-Leibler divergence [2].

The statistical accuracy of the MDL estimator of the density is also bounded by this index of resolvability [2]. Indeed, $Ed^2(p,\hat{p}) \le O(R_n(p))$ where $d(p,q)$ is the Hellinger distance.

Usual choices of parameter discretization lead to a penalty terms of $(m_k/2)\log n$ where m_k is the dimension of the kth family. Then the resolvability is minimax optimal for p in any of the parametric families, but it is suboptimal by a logarithmic factor for p in smooth nonparametric classes.

Here the discretized parameter spaces are modified to allow penalty terms of order m_k, without excessive loss in log likelihood for smooth densities. As a consequence of the removal of the logarithmic factor, the redundancy and the statistical risk will achieve the minimax optimal rates. Other modifications are needed for irregular models such as neural nets, which retain the logarithmic factor.

II. Geometrically regular families

We consider cases in which sequences of parametric models provide accurate approximations with parameter values in an ellipse $E_{r,s,m} = \{\theta \in R^m : \sum_{i=1}^m i^{2s}\theta_i^2 \le r^2\}$ with accuracy $D(p||p_{m,\theta}) \le cr^2/m^{2s}$, where r, s are unknown. The models are chosen such that $D(p_{m,\theta}||p_{m,\theta'})$ is bounded by a constant times the squared Euclidean distance between parameters θ and θ'. These conditions hold for instance when the logarithm of the density on an interval is parameterized using a polynomial or trigonometric expansion of degree m and the true log-density has a bound on the L^2 norm of its sth derivative. (The conditions also hold in a regression setting with Gaussian errors and smooth regression functions modeled using polynomial or trigonometric series.)

The discretized parameter space $\tilde{\Theta}_m$ is taken to be the union for all positive integers r, s, ℓ, of the ellipses $E_{r,s,m}$ intersected with a cubical grid $G_{\ell,m}$ spaced at width $1/\ell$ in each coordinate. An evaluation of the cardinality of the ellipse restricted to the grid shows that we may set $L(\theta|m) = m\log(Jc_r) + O(\log(rs\ell))$, where $J = \ell/m^{s+1/2}$. This codelength is of order m for bounded J whereas it is of order $(m/2)\log n$ when J is of order $\sqrt{n}$. The corresponding MDL criterion leads to estimates $\hat{m}$, $\hat{r}$, $\hat{s}$, $\hat{\ell}$, $\hat{\theta}$, and $\hat{p} = p_{\hat{m},\hat{\theta}}$. Plugging the approximation and codelength bounds into the resolvability leads to the rate $(1/n)^{2s/(2s+1)}$, which is minimax optimal for the redundancy and for the statistical risk. The optimal rate is achieved adaptively, that is, in the absense of knowledge of the index of the smoothness class.

III. Neural nets

Analogous treatment for functions of d variables with the usual expansions leads to an exponentially large parameter dimension $m_k = k^d$ is minimax optimal yet requires exponentially large samples sizes to obtain accurate estimates. For practical inference, it is necessary to consider more restrictive function classes and more parsimonious models.

One useful condition is that the spectral norm $C_f = \int |\omega||\tilde{f}(\omega)|d\omega$ have not too large a value, where $\tilde{f}$ denotes the Fourier transform of the target function f. Sparce trigonometric or sigmoidal expansions $f_m(x) = \sum_{i=1}^m c_i\phi(a_i \cdot x + b_i)$ with a fixed sinusoidal or sigmoidal function ϕ (nonlinearly parameterized by a_i and b_i) provide an approximation error of $||f - f_m||^2 \le C_f^2/m$ and a complexity per sample size of order $md(\log n)/n$, yielding a resolvability of order $c_f(d(\log n)/n)^{1/2}$ as shown in [1]. Here the number and choice of terms is selected using a description length criterion with penalty of (# param.)$\log n$ times a constant. The resulting resolvability exhibits more favorable behaviour in high dimensions than is possible with linear models.

Acknowledgements

The model selection work is joint with Y. Yang and B. Yu.

References

[1] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Mach. Learn.* 14, 115–133, 1994.

[2] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theor.* 37, 1034–1054, 1991.

[3] J. Rissanen, " Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theor.*, 30, 629–636, 1983.

SOME ESTIMATION PROBLEMS IN INFINITE DIMENSIONAL GAUSSIAN WHITE NOISE

I.Ibragimov,[1]St-Peterburg branch of Mathematical Institute RAN,
R.Khasminskii,[2]Dept. of Mathematics Wayne State University,Detroit,MI 48202,USA.

Abstract—Methods of the Information Theory and Approximation Theory are used to obtain the conditions for the existence of consistent estimators for the observations in a Gaussian white noise in a Hilbert space.

0.1 Statement of problem

Let H be a Hilbert space and Q a symmetric positive operator in H.Let $w_Q(t)$ be a Q-Wiener process in the terminology [1]. Let $\mathbf{L}_2(0,1) = \mathbf{L}_2$ be the Hilbert space of H-valued functions s with the inner product and norm

$$(s_1, s_2) = \int_0^1 (s_1(t), s_2(t))_H dt; \|s\|^2 = (s, s).$$

We assume that the process $X_\varepsilon(t), 0 \leq t \leq 1$, is observed, and

$$dX_\varepsilon(t) = S(t)dt + \varepsilon dw_Q(t) \qquad (1)$$

It is known a priori that the "signal" s runs a known set $\Sigma \subseteq \mathbf{L}_2$ and the intensity ε and correlation operator Q of a "noise" $dw_Q(t)$ are known to a statistician. The problem is to estimate the value $\Phi(s)$ of a known function $\Phi : \mathbf{L}_2 \longrightarrow U$ (U is an Euclidean or Hilbert space).The estimation of S and the estimation of finite dimensional parameter in S can be imbedded in this general scheme.

0.2 LAN property

Let $P_S^{(\varepsilon)}$ be the probability distributions associated with X_ε, and P_0^εbe the distribution of $\varepsilon w_Q(t)$.It is well known [2], that for $\Sigma \subseteq Q^{1/2}\mathbf{L}_2$ the measures

[1]Research of this author was supported in part by ISFGrant R36000,Russian Nat.foundation Grant,ONR Grant N00014-93-1-0936.

[2]Research of this author was supported in part by ONR Grant N00014-93-1-0936.

$P_S^{(\varepsilon)}$ are mutually absolutely continuous and

$$\frac{dP^{(\varepsilon)}_{S+\varepsilon Q^{1/2}h}}{dP_S^{(\varepsilon)}}(X_\varepsilon) = \exp\{\int_0^1 (Q^{-1/2}h, dw_Q(t)) - \frac{1}{2}\|h\|^2\}$$

It follows that the family of measures $\{P_S^\varepsilon, S \in \Sigma\}$ satisfies the LAN condition in the sense of [3.]
This fact implies the minimax lower bound of the estimation risks for $\Phi(S)$ and allows to investigate the concept of efficient (asymptotically) estimation for this model.Some natural examples are considered.

0.3 The existence of consistent estimators

If neither $\Phi'(S)$ nor $Q^{1/2}$ are Hilbert-Shmidt operators it is impossible to guarantee the existence even of consistent estimators for $\Phi(S)$.Nevertheless methods of the information theory and theary of approximation allow to propose some necessary and sufficient conditions for the existence of consistent in some metric estimators and to find the rate of convergence of risks to zero when $\varepsilon \to 0$.For example let $Q^{-1/2}\Sigma$ be a bounded set in $\mathbf{L}_2$.Let $\Phi : \mathbf{L}_2 \to B$ be a linear operator,B is a Banach space.Then uniformly consistent estimators of S exist iff $\Phi Q^{1/2}$ is a compact operator.

Our approach generalizes the results of [4].

References

[1] G.Da Prato,J.Zabczik, Stochastic equations in infinite dimensions,Cambridge Univ.Press,1992.

[2] I.Gikhman,A.Skorokhod,The theory of stochastic processes,vol.1,Springer,1974.

[3] I.Ibragimov,R.Khasminsakii,Asymptotically normal families of distributions and efficient estimation,Ann.Stat.,1991,19,4.

[4] I.Ibragimov,R.Khasminskii,On estimation of infinite dimensional parameter in Gaussian white noise,Soviet Math.Doklady,1977,v.236,5.

Local Polynomial Estimation of Regression Functions for Mixing Processes

Elias Masry[1] and Jianqing Fan

Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla,CA 92093-0407
Department of Statistics, University of North Carolina, Chapel Hill, N.C. 27599-3260

Abstract — **Local polynomial fitting for the estimation of a general regression function and its derivatives for ρ-mixing and strongly mixing processes is considered. Joint asymptotic normality for the regression function and its derivatives is established.**

I. Introduction

Local polynomial fitting has been studied in recent years under the assumption of i.i.d. observations and has been shown to possess very useful statistical properties in the context of curve estimation. This paper considers a time series setting and treats the following regression estimation problem. Let $\{X_i, Y_i\}$ be a stationary process and let ψ be a measurable function on the real line. Assume that $E|\psi(Y_1)| < \infty$ and define the regression function

$$m(x) = E[\psi(Y_1)|X_1 = x].$$

Estimates of $m(x)$ and its first p derivatives, via a local polynomials fit, are considered. Special cases include the estimation of conditional distributions and densities $\psi(Y) = I\{Y \le y\}$, conditional moments $\psi(Y) = Y^q$, and d-step prediction in time series $Y_i = X_{i+d}$. The joint asymptotic normality of $m(x)$ and its associated first p derivatives is established for mixing processes $\{X_i, Y_i\}$.

II. Formulation

If the $(p+1)^{th}$ derivative of $m(z)$ at the point x exists, we approximate $m(z)$ locally by a polynomial of order p:

$$m(z) \approx m(x) + \cdots + m^{(p)}(x)(z-x)^p/p! \equiv \beta_0 + \cdots + \beta_p(z-x)^p. \tag{1}$$

One then carries a local polynomial regression by minimizing

$$\sum_{i=1}^{n}\left(\psi(Y_i) - \sum_{j=0}^{p}\beta_j(X_i - x)^j\right)^2 K\left(\frac{X_i - x}{h}\right), \tag{2}$$

where $K(\cdot)$ denotes a nonnegative weight function and h — a smoothing parameter — determines the size of the neighborhood of x. If $\underline{\hat\beta} = (\hat\beta_0, \cdots, \hat\beta_p)$ denotes the solution to the above weighted least squares problem, then by (1), $j!\hat\beta_j(x)$ estimates $m^{(j)}(x), j = 0, \cdots, p$. Minimizing (2) leads to the following set of equations: Let $K_h(x) = K(x/h)/h$ and let

$$s_{n,j} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - x}{h}\right)^j K_h(X_i - x), \tag{3}$$

$$t_{n,j} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - x}{h}\right)^j K_h(X_i - x)\psi(Y_i). \tag{4}$$

[1]This work was supported by the Office of Naval Research under Grant N00014-90-J-1175.

Putting

$$S_n = \begin{pmatrix} s_{n,0} & \cdots & s_{n,p} \\ \vdots & \ddots & \vdots \\ s_{n,p} & \cdots & s_{n,2p} \end{pmatrix}, \quad \underline{t}_n = \begin{pmatrix} t_{n,0} \\ \vdots \\ t_{n,p} \end{pmatrix}, \tag{5}$$

the solution to (2) can be expressed as

$$\underline{\hat\beta}(x) = \mathrm{diag}(1, h^{-1}, \cdots, h^{-p})S_n^{-1}\underline{t}_n. \tag{6}$$

III. Results

Denote

$$\mu_j = \int_{-\infty}^{+\infty} u^j K(u)du, \quad \nu_j = \int_{-\infty}^{+\infty} u^j K^2(u)du.$$

and

$$S = \begin{pmatrix} \mu_0 & \cdots & \mu_p \\ \vdots & \ddots & \vdots \\ \mu_p & \cdots & \mu_{2p} \end{pmatrix}, \quad \tilde S = \begin{pmatrix} \nu_0 & \cdots & \nu_p \\ \vdots & \ddots & \vdots \\ \nu_p & \cdots & \nu_{2p} \end{pmatrix}, \tag{7}$$

$$\underline{\mu} = \begin{pmatrix} \mu_{p+1} \\ \vdots \\ \mu_{2p+1} \end{pmatrix}. \tag{8}$$

We only state here one result along with the conditions on the mixing coefficients. See [1] for the complete analysis and results.

Condition 1. Assume that $h_n \to 0$ and $(nh_n)/\log^2(n) \to \infty$ and put $s_n = (nh_n)^{1/2}/\log n$. For ρ-mixing and strongly mixing processes, we assume that

$$(n/h_n)^{1/2}\rho(s_n) \to 0 \quad \text{and} \quad (n/h_n)^{1/2}\alpha(s_n) \to 0, \text{ as } n \to \infty.$$

Theorem . *Under Condition 1, if $h_n = O(n^{1/(2p+3)})$, then, as $n \to \infty$,*

$$\sqrt{nh_n}\left(\mathrm{diag}(1, \cdots, h_n^p)[\underline{\hat\beta}(x) - \underline{\beta}(x)] - \frac{h_n^{p+1}m^{(p+1)}(x)}{(p+1)!}S^{-1}\underline{\mu}\right) \xrightarrow{\mathcal{L}} N(0, \sigma^2(x)S^{-1}\tilde S S^{-1}/f(x))$$

where $\sigma^2(x) = \mathrm{var}(\psi(Y)|X = x)$, at continuity points of $\sigma^2 f$ whenever $f(x) > 0$.

Remark. The theorem gives the joint asymptotic normality for the estimators $\{\hat m^{(j)}(x) = j!\hat\beta_j(x)\}_{j=0}^p$. The asymptotic normality, "bias", and "variance" of the individual components follow immediately from the theorem.

References

[1] E. Masry and J. Fan " Local polynomial estimation of regression functions for mixing processes." Submitted for publication, November 1993.

The asymptotic normality of global errors for a histogram based density estimate

László Györfi
Technical University of Budapest

Let $\{X_i\}$ be a sequence of i.i.d. real valued random variables with common unknown density f. We denote by μ the measure with density f. We consider the histogram estimate f_n of f built from a partition $\mathcal{P}_n = \{A_{nj}\}$ with interval size $h_n > 0$ that is $f_n(x) = \mu_n(A_n(x))/h_n$, where $A_n(x) = A_{n,i}$ if $x \in A_{n,i}$ and μ_n is the empirical measure. Introduce the following notation:

$$V(\alpha) = Var\left(|N| + \frac{\alpha}{2}[(1 - \frac{|N|}{\alpha})^+]^2\right)$$

where $\alpha > 0$ and N is a standard normal $\mathcal{N}(0,1)$ random variable.

Theorem 1 *([2]): If f is continuously differentiable and if $h_n = cn^{-1/3}$ then*

$$\sqrt{n}\left(\|f_n - f\| - E\|f_n - f\|\right)/\sigma \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $\sigma^2 = \int V\left(\frac{c^{3/2}|f'|}{2\sqrt{f}}\right) f.$

One can show that $\sigma^2 \leq 1 - \frac{2}{\pi}$. This should be compared to the rate of convergence of $E\|f_n - f\|$, which is at least of order $n^{-1/3}$ for differentiable f, and it can be achieved for $h_n = cn^{-1/3}$.

We consider the problem of estimating an unknown probability density function in information divergence. If μ and ν are probability measures on the real line, absolutely continuous with respect to a σ-finite measure λ with densities f and g respectively, then the information divergence between μ and ν is defined by

$$I(\mu,\nu) = \int_R f(x)\log\frac{f(x)}{g(x)}\lambda(dx) = D(f,g).$$

Barron, Györfi and van der Meulen (1992) showed that if there exists a known density g such that $D(f,g) < \infty$, then one can construct a density estimator as follows: define a sequence of integers m_n, and put $h_n = 1/m_n$. Let ν denote the probability measure with density g. Introduce partitions $P_n = \{A_{n,1}, A_{n,2}, ..., A_{n,m_n}\}$, $n = 2, ...,$ of the real line such that the $A_{n,i}$'s are intervals with $\nu(A_{n,i}) = h_n$. For a given sequence $a_n = 1/(nh_n+1)$ consider the following density estimate:

$$f_n(x) = ((1 - a_n)\mu_n(A_n(x))/h_n + a_n)g(x).$$

If in addition $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n = \infty$, then $\lim_{n\to\infty} D(f,f_n) = 0$ a.s.

Theorem 2 *([3]): Let S_μ be the support set of μ. Under the conditions of consistency*

$$n\sqrt{2h_n}[D(f,f_n) - E(D(f,f_n))] \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2),$$

where $\sigma^2 = \nu(S_\mu) > 0$.

For the choice $m_n = n^{1/3}$ the rate of convergence of the random part of the divergence error is of order $n^{-5/6}$, and under some restrictive conditions on the unknown density f

$$E(D(f,f_n)) \leq O(n^{-2/3}).$$

References

[1] Barron, A. R., Györfi, L. and van der Meulen, E. C. (1992) "Distribution estimates consistent in total variation and in two types of information divergence". *IEEE Trans. on Information Theory*, 38, pp. 1437-1454.

[2] Berlinet, A., Devroye L. and Györfi, L. (1994) "The asymptotic normality of L_1 error in density estimation" (submitted for publication)

[3] Berlinet, A., Györfi, L. and van der Meulen, E. C. (1994) "The asymptotic normality of information divergence error for a histogram based density estimate" (submitted for publication)

Bandwidth Choice and Convergence Rates in Density Estimation with Long-Range Dependent Data

Peter Hall, Soumendra Nath Lahiri, Young K. Truong
Centre for Mathematics and its Applications, Australian National University,
Canberra, ACT 0200, Australia

Abstract — **We discuss optimal bandwidth choice and optimal convergence rates for density estimation with dependent data, as the amount of information in the sample is altered by adjusting the range of dependence.**

I. Introduction

Assume that data are observed from a stationary stochastic process that may be taken to be an unknown function of a Gaussian process. Thus, the strength of dependence is determined entirely by a single sequence of numbers, the covariances $\gamma(i)$; this makes it relatively straightforward to appreciate the influence of different strengths of dependence on various aspects of bandwidth choice, even up to terms of second or third order. Let us assume here, for the sake of simplicity, that $\gamma(i) \sim ci^{-\alpha}$ for constants $c \neq 0$ and $\alpha > 0$. Then smaller values of α correspond to less information in a data sequence of given length n from the process, and hence to slower convergence rates. Surprisingly, the traditional dichotomy of short-range versus long-range dependence, or equivalently $\alpha > 1$ versus $\alpha \leq 1$, does not have a major role to play in the bandwidth choice problem. We shall discuss the effect that the value of α has on optimal bandwidth choice and, correspondingly, on convergence rates.

II. Outline of Main Results

In the case of density estimation based on a second-order kernel, the "barrier" normally encountered at $\alpha = 1$ occurs instead at $\alpha = 4/5$. When $\alpha > 4/5$ the minimum mean integrated squared error (MISE) is asymptotic to a constant multiple of $n^{-4/5}$, which is identical to that value which it enjoys in the case of independence (effectively, $\alpha = \infty$). Indeed, even the value of the constant is identical to that which it would be for independent data. Furthermore, for such α's the deterministic bandwidth that minimizes MISE agrees even to *second* order with its counterpart in the case of independent data; under short-range dependence, where $\alpha > 1$, the agreement is up to (but not including) *third* order.

When $\alpha \leq 4/5$, minimum MISE is of size $n^{-\alpha}$, but curiously, provided that $2/5 < \alpha \leq \infty$, the bandwidth that produces the overall minimum still agrees to first order with its counterpart in the case of independent data. Thus, very-long-range dependence is allowable before much change has to be made to the optimal bandwidth formula. Only when $\alpha \leq 2/5$, which is a context of particularly long-range dependence, is there a large difference between the first-order properties of the MISE-optimal bandwidth under dependence, and its counterpart for independent data.

What is more, even when $\alpha \leq 2/5$ the bandwidth appropriate for independent data produces first-order minimization of MISE. This is a consequence of the fact that, whenever $\alpha < 4/5$, adjusting the bandwidth in the vicinity of the optimum has an effect only on second- and higher-order terms; to first order, MISE does not depend on bandwidth. This result is rather striking to researchers who are familiar only with the case of independent data, where first-order adjustments to bandwidth always affect first-order features of performance.

More generally, if the kernel is of order $r \geq 2$ then the "boundaries" at $2/5$ and $4/5$ change to $r/(2r+1)$ and $2r/(2r+1)$, respectively.

These results indicate that those practical bandwidth-choice rules that have been proposed for independent data and are based on plug-in rules, have straightforward generalizations to certain types of dependent data, even under very long-range dependence. Generally speaking this is true, although there are some qualifications.

SESSION V

Markov Random Fields

Large deviations and consistent estimates for Gibbs Random fields

Francis COMETS[1]
Univ. Paris 7 - Denis Diderot, Mathematiques case 7012, 2 place Jussieu 75251 Paris cedex 05, France

Abstract **- Large deviations estimates yield a convenient tool to study asymptotics of Gibbs fields. Applications to parametric estimation and detection of phase transition are given.**

I. INTRODUCTION

Gibbs Random Fields (GRF) provide pertinent statistical models for spacial data X_i, $i \in Z^d$, where important features of the dependence structure can be captured in a very natural way. An important issue is image analysis via D. § S. Geman's Bayesian approach.

II. PARAMETRIC FAMILIES OF GRF

We are given for each $\theta \in \Theta \subset R^p$ a compatible family indexed by finite $\Lambda \subset Z^d$ of conditional distributions $\Pi_{\theta,\Lambda}$ of $X_\Lambda = (X_i)_{i \in \Lambda}$ given X_{Λ^c}; these $\Pi_{\theta,\Lambda}$ are related by a natural translation invariance property. A distribution P having these specified conditional distributions $\Pi_{\theta,\Lambda}$ is called a GRF. First order phase transition occurs when the set $G(\theta)$ of such GRF contains more than one element; this situation is characterized by the fact that the set $G_s(\theta)$ of stationary elements of $G(\theta)$ does not reduce to a singleton. Note that this is not an effect of ill parametrization, but an intrinsic phenomenon. Then, some GRF are not ergodic, and even worse, it may exist some GRF which are not translation invariant. Statisticians in front of real data should not assume invariance in general.

III. LARGE DEVIATIONS

The empirical field based on a configuration $x = (x_i)_{i \in Z^d}$ and on a cubic box Λ is

$$R_{\Lambda,x} = \frac{1}{|\Lambda|} \sum_{i \in \Lambda} \delta_{\tau_i x}$$

with τ_i the shift operator. When it holds the ergodic theorem states that the empirical field is a good guess for the actual GRF. In general we will use the following as a substitute. In general, large deviations estimates hold, for all $P \in G(\theta)$, and they mean heuristically

$$\text{" } P\{R_{\Lambda,X} \text{ close to } Q\} \sim \exp - |\Lambda| \, I_\theta(Q) \text{ "} \qquad (1)$$

for some non-negative entropy functional I_θ defined on the set of random fields. Moreover

$$I_\theta(Q) = 0 \Leftrightarrow Q \in G_s(\theta) \qquad (2)$$

The relations (1) and (2) simply state that the data will not behave worse than the worse stationary GRF.

IV. CONSISTENCY CRITERIA FOR PARAMETRIC ESTIMATORS

Let Λ be the window of observation, and $\widehat{\theta_\Lambda}$ be any maximizer of some objective function $\theta \to k_\Lambda(\theta ; X_\Lambda)$. Assume that there exists a real continuous function $K(\theta ; Q)$ with

i) $k_\Lambda(\theta ; x_\Lambda) = K(\theta ; R_{\Lambda,x}) + \varepsilon_\Lambda$ and, $\lim_{\Lambda \to Z^d} \sup_{x, \theta \in \Theta} \varepsilon_\Lambda = 0$

ii) θ is the unique maximizer of $K(. ; P)$, $\theta \in \Theta$, $P \in G_s(\theta)$

iii) Θ is compact.

Then, $\widehat{\theta_\Lambda}$ is a.s. consistent.

The previous criterium applies to classical estimators in a general setup. The question of asymptotic optimality also suffers from the breakdown in the central limit theorem, related to phase transition : it can be treated via large deviation using Bahadur's approach. At last, prior to the use of gaussian statistics and tests, one would like to know from the data themselves if phase transition hold or not.

V. DETECTING PHASE TRANSITION

For cubic boxes Λ we now choose smaller cubic boxes Λ' such that, as $\Lambda \uparrow Z^d$,

$$|\Lambda'|^{-1} \text{Log} \, |\Lambda| \to 0 , \quad |\Lambda'|^{-1} (\text{Log} \, |\Lambda|)^{d/(d-1)} \to \infty \qquad (3)$$

Define the set $\Delta_\Lambda(X)$ of moving empirical fields $R_{i+\Lambda',X}$ based on all the translates $i + \Lambda'$ of Λ' which are included in the window of observation Λ. Generalized Erdös-Rényi laws state that for $P \in G(\theta)$, under the condition (3) it holds P - a.s.

$$\lim_{\Lambda \to Z^d} \Delta_\Lambda(X) = G_s(\theta) \qquad (4)$$

in the sense of Hausdorff convergence of closed sets.

The result (4) shows how to estimate consistently the set of all stationary GRF with the same parameter θ as the underlying one. Then one can asymptotically detect phase transition from a single sample. More practical versions of (4) may be given, and studied via simulation experiments.

[1] URA CNRS 1321 "Statistique et Modèles Aléatoires"

Large deviations and the rate distortion theorem for Gibbs distributions

Yali Amit[1]

Department of Statistics, University of Chicago, Chicago IL, 60637 USA

Abstract — **Large deviation theory is used to obtain the rate distortion theorem for Gibbs distributions together with exponentially small error probabilities.**

Let σ be a Gibbs distribution on $\Omega_0^{Z^2}$ with finite range interaction and Ω_0 finite. For simplicity we assume that σ is unique. For any domain $G \subset Z^2$ containing the origin, define $R_{\Lambda,G}(\omega) = \sum_{x\in\Lambda} \delta_{\omega_{G+x}}$, where $G \subset\subset \Lambda \subset Z^2$. Large deviation theorems [1] provide asymptotically exponential upper and lower bounds on the probability that the empirical distribution $R_{\Lambda,G}(\omega)$ under σ, deviates in variational norm from the marginal σ_G of σ on G, as Λ tends to infinity. In particular these hold if σ is a product measure. Using these theorems many of the standard asymptotic results of errorless coding theory can be neatly formulated and extended to Gibbs random fields, see [2].

Here we present the application of these theorems to coding with distortion, more or less following the proof in [3]. Let $l(\cdot,\cdot)$ be some distance on Ω_0 and define

$$\lambda(\omega_\Lambda, \omega'_\Lambda) = \frac{1}{|\Lambda|}\sum_{x\in\Lambda} l(\omega_x, \omega'_x).$$

Given $\delta > 0$ and $\lambda > 0$, what size codebook is needed so that with very high probability a random sample ω_Λ from σ will find a code word ω'_Λ such that $\lambda(\omega_\Lambda, \omega'_L) < \lambda + \delta$?

Let Λ_n be an increasing sequence of $n \times n$ domains . Fix l and set $k_n = n/l$. Henceforth the n subscript is omitted for notational ease. Let $G_{ij}, i,j = 0,\dots k-1$ be the k^2 non-overlapping $l \times l$ domains in Λ. Set $G = G_{00}$. Let $Q(\omega'_G|\omega_G)$ be a conditional probability distribution, and $Q(\omega_G, \omega'_G) = \sigma_G(\omega_G)Q(\omega'_G|\omega_G)$ be the joint probability on $\Omega_0^G \times \Omega_0^G$. The marginal on the second coordinate is denoted $Q_2(\omega'_G)$. Define

$$\begin{aligned}\lambda_Q &= \sum_{\omega_G,\omega'_G} \lambda(\omega_G,\omega'_G)Q(\omega_G,\omega'_G)\\ R_G(Q) &= \frac{1}{|G|}\sum_{\omega_G,\omega'_G} Q(\omega_G,\omega'_G)\log\frac{Q(\omega'_G|\omega_G)}{Q_2(\omega'_G)}\end{aligned}$$

Thus λ_Q is the expected distortion under the joint distribution, and $R_G(Q)$ is the average mutual information of the two coordinates.

For any two distributions σ_G, π_G on Ω_0^G let $|\sigma_G - \pi_G| = \max_{\Omega_0^G} |\sigma_G(\omega_G) - \pi_G(\omega_G)|$. Let $\tilde{R}_{\Lambda,G}(\omega)$ be the block empirical distribution on Ω_0^G, considering only disjoint blocks. Applying the large deviation theorems to block Gibbs distributions where each disjoint G block is aggregated as one site we get that outside a set $B_{\epsilon,\Lambda}$ of exponentially small probability in n^2, the frequency of occurence $n(\omega_G)$ of a specific configuration ω_G, without overlaps, in ω_Λ is within ϵ of its underlying probability, $\sigma_G(\omega_G)$.

For $u_G \in \Omega_0^G$ let $c(u_G) = \sum_{\omega_G} \lambda(u_G,\omega_G)Q(u_G,\omega_G)/\lambda_Q$. For each domain G_{ij} choose $\omega'_{G_{ij}}$ *independently* from the distribution Q_2, to obtain another configuration ω'_Λ from the product distribution $\mathcal{Q} = \otimes_{i,j=1}^k Q_2(\cdot)$ on Ω^Λ. Let ω_Λ be a configuration in $B^c_{\epsilon,\Lambda}$. Using the fact that $|n(u_G)/k^2 - \sigma_G(u_G)| < \epsilon$, the probability that $\lambda(\omega_\Lambda,\omega'_\Lambda) \le \lambda_Q + \delta$ is bounded below by

$$\prod_{n(u_G)>0} Q_2^{n(u_G)}\left(\frac{1}{n(u_G)}\sum_{\omega_{G_{ij}}=u_G} \lambda(u_{G_{ij}},\omega'_{G_{ij}}) \le \beta(u_G)\right),$$

where $\beta(u_G) = c(u_G)(\lambda_Q + \delta/2)/(\sigma_G(u_G) - \epsilon)$.

Using large deviation results for i.i.d distributions, given arbitrary $\gamma > 0$, for sufficiently large k, each term in the above product is bounded below by $\exp[-k^2(\sigma_G(u_G)+\epsilon)(J(u_G)+\gamma)]$, where $J(u_G) = \inf_{F(u_G)} D(\pi_G, Q_2)$ is the infimum ofRKL divergences with respect to Q_2 over the set of measures

$$F(u_G) = \left\{\pi_G; \int \lambda(u_G,\omega'_G)\pi_G(d\omega'_G) \le \frac{c(u_G)(\lambda_Q+\delta/2)}{\sigma_G(u_G)-\epsilon}\right\}.$$

From the choice of $c(u_G)$ it follows that the conditional distribution $Q(\cdot|u_G) \in F(u_G)$. Aggregating this over all u_G's found in ω_Λ we have

$$\underline{\lim}\frac{1}{n^2}\log\mathcal{Q}\left(\lambda(\omega_\Lambda,\omega'_\Lambda) \le \lambda_Q + \delta\right) \ge R_G(Q) - \gamma',$$

with $\gamma' \to 0$ as $\epsilon \to 0$.

Taking $L = \exp\left[n^2(R_G(Q)+\gamma'+\gamma'')\right]$ with $\gamma'' > 0$, choose L independent samples $v^\alpha_\Lambda, \alpha = 1,\dots,L$, from $\mathcal{Q}$. Using the lower bound above, it is easily shown that with probability exponentially close to 1, *every* element of $B^c_{\epsilon,\Lambda}$ is within distance less than $\lambda_Q + \delta$ from at least one of the v^α_Λ in the random sample, and $\gamma'' \to 0$ as $\epsilon \to 0$.

Theorem: There exist constants $c, d, \alpha, \beta > 0$ such that with probability $1 - ce^{-n^2\alpha}$ a random choice of L independent samples from the distribution $\mathcal{Q}$ on Ω^Λ will provide a codebook of rate $(R_G(Q)+\gamma'+\gamma'')$ per pixel, for which any configuration $\omega_\Lambda \in B^c_{\epsilon,\Lambda}$, has a codeword v_Λ such that $\lambda(\omega_\Lambda, v_\Lambda) < \lambda_Q + \delta$, and $\sigma(B_{\epsilon,\Lambda}) < de^{-n^2\beta}$. Moreover $\gamma', \gamma'' \to 0$ as $\epsilon \to 0$ so that it is asymptotically possible to code samples from σ with minimal rate

$$R_G(\lambda) = \inf_Q\{R_G(Q); \lambda_Q \le \lambda\}\ \bullet$$

Observe that this rate distortion curve depends on the base domain G. The larger G the lower the rates will be for fixed distortion.

[1]This work was supported by Grant ARO DAAL03-92-G-0322.

References

[1] Comets F., "Grandes deviations pour les champs de Gibbs sur Z^{d}", *C.R. Acad. Sc., Paris*, t. 303, Serie I, no. 11. 1986

[2] Amit Y. and Miller M., "Large deviations for coding Markov chains and Gibbs random fields" *IEEE IT*, vol. 39, no. 1, pp. 109-118. 1993

[3] Berger T., *Rate Distortion Theory*, Prentice Hall, 1971

Estimation and Prediction for (Mostly Gaussian) Markov Fields in the Continuum

Loren D. Pitt[1]
Mathematics Dept., Univ. of Virginia, Charlottesville, VA 22903, USA

Abstract — **We present a survey of design problems and results that arise in the prediction and parameter estimation of stochastic partial differential equations. The aim is to better understand some unavoidable errors that occur in the discretization of SPDEs, and available methods for minimizing these errors.**

I. Introduction

Solutions to many prediction and estimation problems associated with continuous Gaussian Markov fields satisfy minimum principles and may be characterized as solutions of stochastic boundary value or initial value problems, see e.g. [1], [2], [3], and [4]. These characterizations provide a theoretical basis for the calculation, but in the implementation of these calculations numerous issues arise. A typical problem may involve a smooth elliptic boundary problem on a smooth domain, with boundary data that must be empirically determined, but typically this data will be generalized functions and can not be interpreted as classical functions. A careful analysis is required to determine the relative merits and limitations of different discretizations of such a problem. This paper presents examples which illustrate these issues, and where the required analysis has, at least in part, been completed.

II. A General Design Problem with an Illustrative Example

Consider a random field $\{\phi(t,x) : t \in R, x \in R^d\}$ that cannot be observed on a restricted set $D \subset R \times R^d$. It is desired to observe ϕ off the set D, and, based on these observations, to calculate the conditional expectation $\hat{\phi}_D(t,x)$ for $(t,x) \in D$. Of course, in fact, ϕ can only be observed on a finite set of N times and places $(t_j, x_j) \in D$, and N may be very limited. The following questions arise in considering the merits of designs and computational recopies. If N is fixed, what is the smallest possible prediction error $e^2(x,N)$, and what is the limiting value $e^2(x,\infty)$? What are the asymptotics of

$$e^2(x,\infty) - e^2(x,N)$$

and of

$$\frac{e^2(x,\infty) - e^2(x,N)}{e^2(x,N)}?$$

Where should the N sites $\{(t_j,x_j)\}$ be located to approximately achieve the minimum error $e^2(x,N)$?

The simplest illustrative special case for these problems occurs in the time independent $d = 2$ case when ϕ satisfies the elliptic SPDE

$$\phi(x) - \Delta\phi(x) = \dot{w}(x),$$

with $\dot{w}(x)$, a Gaussian white noise in the plane. In this case, when $D \subset R^2$ is a bounded domain with smooth boundary Γ, typical results are

A.

$$(I - \Delta)^2 \hat{\phi}_D(x) = 0$$

for all $x \in D$, and $\hat{\phi}_D$ satisfies the boundary conditions $\hat{\phi}_D(x) = \phi(x)$ and $\partial_n \hat{\phi}_D(x) = \partial_n \phi(x)$ on Γ.

B.

$$e^2(x,\infty) = G_D(x,x),$$

where $G_D(x,y)$ is the Green's function for $(I - \Delta)^2$ on the domain D.

A careful analysis of the errors made in discretizing the Poisson integral representation of $\hat{\phi}_D$ in A yields, see [4], [5],

C.

$$e^2(x,\infty) - e^2(x,N) = 1/N,$$

together with precise numerical constants and asymptotically optimal locations for the sites $\{x_j\}$.

[1]Supported ONR Contract No. N00014-90-J-1639.

References

[1] M. Hübner, R. Khasminskii, B. L. Rozovskii, "Two examples of parameter estimation for stochastic partial differential equations," preprint to appear in *Theory of Prob. and Rel. Fields*, 1995

[2] L. I. Piterbarg, "On prediction of a class of random fields," *Theory Prob. Appl.*, vol. 28, pp. 184-191, 1983

[3] L. D. Pitt, "A Markov property for Gaussian processes with a multidimensional parameter," *Arch. Rat. Mech. Anal.*, vol. 43, pp. 367-391, 1971

[4] L. D. Pitt, D. Y. Wong, "On stochastic Elliptic boundary value problems associated with Gaussian Markov random fields," in *Stochastic Partial Differential Equations and Their Applications*, B. L. Rozovskii and R. B. Sowers (Eds.), Lecture Notes in Control and Information Sciences 176, Springer, New York, pp. 222-237, 1992

[5] L. D. Pitt, R. Robeva, and D. Y. Wong, "An error analysis for the numerical calculation of certain random integrals: part 1," preprint to appear, *Ann. Appl. Prob.* 1995

Markov chain Monte Carlo algorithms

by

Jeffrey S. Rosenthal*

Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1

Phone: (416) 978-4594. Internet: jeff@utstat.toronto.edu

Abstract. We briefly describe Markov chain Monte Carlo algorithms, such as the Gibbs Sampler and the Metropolis-Hastings Algorithm, which are frequently used in the statistics literature to explore complicated probability distributions. We present a general method for proving rigorous, *a priori* bounds on the number of iterations required to achieve convergence of the algorithms.

I. Introduction.

Markov chain Monte Carlo techniques have become very popular in recent years as a way of generating a sample from complicated probability distributions (such as posterior distributions in Bayesian inference problems). The idea of such algorithms is to define a Markov chain which has as its stationary distribution, the distribution $\pi(\cdot)$ of interest.

Procedures for defining the Markov chain include the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), whereby the Markov chain proceeds by "proposing" a new point according to some scheme, and then "accepting" that point with a certain probability, chosen to make the Markov chain reversible with respect to $\pi(\cdot)$; and the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990), whereby the Markov chain proceeds by updating the various coordinates of the point in turn according to the correct *conditional distribution* as indicated by $\pi(\cdot)$.

A fundamental issue regarding such techniques is their convergence properties, specifically whether or not the algorithm will converge to the correct distribution, and if so how quickly.

II. A quantitative convergence result.

We describe here a general method (Rosenthal, 1993, Theorem 12) for proving quantitative bounds on the time to stationarity of a Markov chain. The method requires only that we verify a drift condition and a minorization condition, for the Markov chain of interest. In certain simple cases, the bound appears to be small enough to be of practical use; see Rosenthal (1993, 1994) and references therein. For related results see Meyn and Tweedie (1993).

* Supported in part by NSERC of Canada.

Proposition. *Let $P(x, \cdot)$ be the transition probabilities for a Markov chain with stationary distribution $\pi(\cdot)$. Suppose there exist $\epsilon > 0$, $0 < \lambda < 1$, $0 < \Lambda < \infty$, $d > \frac{2\Lambda}{1-\lambda}$, $f : \mathcal{X} \to \mathbf{R}^{\geq 0}$, and a probability measure $Q(\cdot)$ on $\mathcal{X}$, such that $\mathbf{E}(f(X_1) \mid X_0 = x) \leq \lambda f(x) + \Lambda$ for $x \in \mathcal{X}$, and $P(x, \cdot) \geq \epsilon Q(\cdot)$ for $x \in f_d$, where $f_d = \{x \in \mathcal{X} \mid f(x) \leq d\}$. Then for any $0 < r < 1$, the total variation distance to the stationary distribution after k iterations is bounded above by*

$$(1-\epsilon)^{rk} + \left(\alpha^{-(1-r)}\gamma^r\right)^k \left(1 + \frac{\Lambda}{1-\lambda} + \mathbf{E}(f(X_0))\right),$$

where $\alpha^{-1} = \frac{1+2\Lambda+\lambda d}{1+d}$, $\gamma = 1 + 2(\lambda d + \Lambda)$.

REFERENCES

A.E. Gelfand and A.F.M. Smith (1990), Sampling based approaches to calculating marginal densities. J. Amer. Stat. Assoc. **85**, 398-409.

S. Geman and D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence **6**, 721-741.

W.K. Hastings (1970), Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97-109.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087-1091.

S.P. Meyn and R.L. Tweedie (1993), Computable bounds for convergence rates of Markov chains. Tech. Rep., Dept. of Statistics, Colorado State University.

J.S. Rosenthal (1993), Minorization conditions and convergence rates for Markov chain Monte Carlo. Tech. Rep. **9321**, Dept. of Statistics, University of Toronto. J. Amer. Stat. Assoc., to appear.

J.S. Rosenthal (1994), Analysis of the Gibbs sampler for a model related to James-Stein estimators. Tech. Rep. **9413**, Dept. of Statistics, University of Toronto.

Markov Random Fields on Graphs for Natural Languages

Joseph A. O'Sullivan, Kevin Mark, and Michael I. Miller
Department of Electrical Engineering, Campus Box 1127, Washington University, St. Louis, MO 63130
(314) 935-4173, e-mail: *jao@ee.wustl.edu*

I. Introduction

The use of model-based methods for data compression for English dates back at least to Shannon's Markov chain (n-gram) models, where the probability of the next word given all previous words equals the probability of the next word given the previous n-1 words. A second approach seeks to model the hierarchical nature of language via tree graph structures arising from a context-free language (CFL). Neither the n-gram nor the CFL models approach the data compression predicted by the entropy of English as estimated by Shannon and Cover and King. This paper presents two recently proposed models that incorporate the benefits of both the n-gram model and the tree-based models [1,2]. In either case the neighborhood structure on the syntactic variables is determined by the tree while the neighborhood structure of the words is determined by the n-gram and the parent syntactic variable (preterminal) in the tree. Having both types of neighbors for the words should yield decreased entropy of the model and hence fewer bits per word in data compression. To motivate estimation of model parameters, some results in estimating parameters for random branching processes is reviewed.

II. Random Branching Processes

A stochastic context-free grammar (SCFG) is a quintuple $< V_N, V_T, \mathbf{R}, \sigma_0, \mathbf{P} >$, where V_N is the set of V syntactic variables σ_k, V_T is the finite set of words or terminals, $\mathbf{R}$ is the finite set of rules, σ_0 is the starting syntactic variable, and $\mathbf{P}$ is the set of conditional probabilities for the rules, conditioned on the syntactic variable being rewritten. The probability of a derivation from the SCFG is the product of all of the probabilities used in the derivation. A tree $\mathbf{T}$ is associated with a derivation by mapping syntactic variables used in the derivation to nodes in the tree; the rule used for rewriting each syntactic variable determines the children nodes. Define the mean matrix $\mathbf{M}$ to have its j,k entry equal to the expected number of σ_k that result from rewriting σ_j. $\mathbf{M}$ has largest eigenvalue ρ greater than or less than one according to whether the SCFG is supercritical or subcritical.

A function of a tree, f, is said to be additive on the rules with atomic function $\bar{f}$ if $f(\mathbf{T}) = \sum \bar{f}(r)$, where the sum is over the rules r used. Let $n(\mathbf{T})$ equal the number of syntactic variables in the tree $\mathbf{T}$. Assume that $\bar{f}$ is finite for all $r \in \mathbf{R}$. Let $\mathbf{T}_K$ be the truncation of $\mathbf{T}$ at derivation depth K.

Theorem 1 [3]: Suppose that $\mathbf{M}$ is strongly connected with largest eigenvalue $\rho > 1$ and associated left eigenvector v. Then for almost all infinite length derivations,

$$\lim_{K\to\infty} \frac{f(\mathbf{T}_K)}{n(\mathbf{T}_K)} = \sum_{i=1}^{V} v(i) \sum_{k=1}^{J_i} p(i,k)\bar{f}(i,k) = v\bar{\mathbf{f}}\,, \qquad (1)$$

where J_i is the number of rules in $\mathbf{R}(i)$, the set of rules for rewriting σ_i; $p(i,k)$ is the probability of that rule; $\bar{\mathbf{f}}$ is the $V \times 1$ vector with ith entry $\sum_{k=1}^{J_i} p(i,k)\bar{f}(i,k)$.

Extensions of this theorem include the convergence of ratios of such functions [3]. Notice that $f(\mathbf{T}_K)$ and $n(\mathbf{T}_K)$ are derivation statistics that can be used to estimate model parameters via (1). The SCFG's used to model English are usually subcritical. The corresponding result requires a sequence of independent derivations from the SCFG.

Theorem 2: Suppose that $\mathbf{M}$ has largest eigenvalue less than one. Let $\{\mathbf{T}^{(m)}\}$ be a sequence of independent trees each having distribution determined by the SCFG. Then

$$\lim_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{T}^{(m)}) = \mathbf{z}_0(\mathbf{I}-\mathbf{M})^{-1}\bar{\mathbf{f}}\,, \qquad (2)$$

where $\mathbf{z}_0$ is the $1 \times V$ unit vector with entry one in the location corresponding to the syntactic variable σ_0.

III. Proposed Language Models

The first proposed language model adds n-gram constraints to the tree-based models. For a given word string (sentence) $W_{1,N} = w_1 w_2 \cdots w_N$, define the relative frequency of $\omega_j\omega_i$ by

$$\frac{n_{\omega_j\omega_i}(W_{1,N})}{N-1} = \frac{1}{N-1} \sum_{k=1}^{N-1} 1_{\omega_j\omega_i}(w_k, w_{k+1})\,. \qquad (3)$$

Theorem 3 [2]: The probability distribution on trees $\mathbf{T}$, p, minimizing the Kullback-Leibler distance from the distribution π defined by the SCFG, $\sum p(\mathbf{T}) \log \frac{p(\mathbf{T})}{\pi(\mathbf{T})}$, subject to the bigram constraints $E[n_{\omega_j\omega_i}(W_{1,N})/(N-1)] = H_{\omega_j\omega_i}$, $\omega_j, \omega_i \in V_T$, is

$$p(\mathbf{T}) = \frac{1}{Z} \exp[\frac{1}{N_T - 1} \sum_{\omega_j \in V_T} \sum_{\omega_i \in V_T} \alpha_{\omega_j\omega_i} n_{\omega_j\omega_i}(W_{1,N_T})]\pi(\mathbf{T})\,, \quad (4)$$

where Z is the normalizing constant and the $\alpha_{\omega_j\omega_i}$ are the Lagrange multipliers chosen to satisfy the constraints.

The distribution (4) induces the neighborhood structure discussed in the introduction. As with many random field models, computing Z is problematic. This motivates a second model, the mixed tree/chain graph. In this model, let $\mathbf{T}$ be the tree down to the preterminal layer, and label the preterminals for a particular derivation by γ_k, $k = 1, 2, \ldots, N_T$. The probabilities of words are determined by conditional probabilities on the words, $p(w_k|w_{k-1}, \gamma_k)$, and the SCFG down to the preterminal level.

Issues that are under investigation include: the decrease in entropy obtained by using successively more complicated models; comparative performance of different models as a function of the number of parameters; estimation of parameters in the two proposed models using the Penn TreeBank; determining the compressibility of the Penn TreeBank using our models.

1. M. I. Miller and K. E. Mark, "Inference on Pattern Theoretic Representations: Applications to Shape and Natural Languages," to appear Proc. IMA Workshop on Image and Speech Models, May 1994, Springer-Verlag.
2. K. E. Mark, M. I. Miller, U. Grenander, and S. Abney, "Parameter estimation for constrained context-free language models," in *Proc. DARPA Speech and Natural Language Workshop*.New York: Morgan Kaufman, 1992.
3. J. A. O'Sullivan and M. I. Miller, "Almost sure convergence for functions of random branching processes," to appear *IEEE Trans. Inform. Theory,* 1995.

Equilibria in Infinite Random Graphs

Bruce Hajek *
Coordinated Science Laboratory and the
Department of Electrical and Computer Engineering
University of Illinois, Urbana, Illinois 61801, USA

Abstract – **A load balancing problem is formulated for infinite networks or graphs. There are overlapping sets of locations, each set having an associated possibly random amount of load to be distributed. The total load at a location is the sum of the contributions due to the sets that contain it. Equilibrium is said to hold if the load corresponding to any one set cannot be reassigned to improve the balance of total loads. The set of possible equilibria, or balanced load vectors, is examined. The balanced load vector is shown to be unique for Euclidean lattice networks, in which the sets correspond to pairs of neighboring nodes in a rectangular lattice in finite dimensions. A method for computing the load distribution is explored for tree networks. An FKG type inequality is proved. The concept of load percolation is introduced and is shown to be associated with infinite sets of locations with identical load.**

SUMMARY

A balancing problem is specified by a collection (U, V, N, m), where U and V are finite or countably infinite sets, $N = \{N(u) : u \in U\}$ where $N(u)$ is a finite subset of V for each $u \in U$, and $m = (m_u : u \in U)$ where $m_u > 0$ for all u. For example, U may denote the edges of a (possibly infinite) graph, V the vertices, and $N(u)$ the set consisting of the two endpoints of edge u for each u. An *assignment* vector is a vector $f = (f_{u,v} : u \in U,\ v \in V)$, with nonnegative entries. It is said to meet the demand m if

$$\sum_{v \in N(u)} f_{u,v} = m_u \quad \text{for } u \in U, \tag{1}$$

The total load at v, $x(v)$, is given by

$$x(v) = \sum_{u \in U} f_{u,v} \quad \text{for } v \in V. \tag{2}$$

A vector $x = (x(v) : v \in V)$ so arising from an assignment vector f meeting the demand is called a *load vector*. A load vector x is said to be *balanced*, if for some corresponding f, the following conditions hold: For all $u \in U$ and all $v, v' \in N(u)$, $f_{u,v} = 0$ whenever $x(v) > x(v')$.

The main questions addressed in this paper can be stated in broad terms as follows. How can the set of balanced load vectors be characterized? It is not difficult to show that balanced load vectors exist, but are they unique? What is the distribution of the load at a given location for a balanced load vector when the demand vector is random? Finally, what "global" or long-range effects can be observed in balanced load vectors?

The highlights of this paper are summarized as follows. The concept of load balancing on an infinite network is introduced (in somewhat more generality than the above). Minimal and maximal balanced load vectors are shown to exist, and the idea of load balancing in finite subsets with boundary conditions is used to exhibit a one-parameter family of balanced load vectors whenever the balanced load vector is not unique. It is shown that the balanced load vector is unique for a wide class of networks including rectangular lattice networks. The concept of τ-surplus is used to characterize the possible distributions of the load at a location in a tree network with independent, identically distributed demands. The case of Bernoulli demand and exponentially distributed demands are investigated in some detail. Finally a notion of long range interaction, load percolation, is introduced. Load percolation is shown to imply the existence of infinite connected sets of locations with identical load.

*This work was supported by JSEP Contract N00014-90-J-1270

SESSION VI

Theory and Applications of Wavelets

Selection of Best Bases for Classification and Regression

Ronald R. Coifman[1] and Naoki Saito[1,2]

[1] Department of Mathematics, Yale University, New Haven, CT 06520

[2] Schlumberger-Doll Research, Old Quarry Rd., Ridgefield, CT 06877

Abstract — **We describe extensions to the "best-basis" method to select orthonormal bases suitable for signal classification (or regression) problems from a collection of orthonormal bases using the relative entropy (or regression errors). Once these bases are selected, the most significant coordinates are fed into a traditional classifier (or regression method) such as Linear Discriminant Analysis (LDA) or a Classification and Regression Tree (CART). The performance of these statistical methods is enhanced since the proposed methods reduce the dimensionality of the problems by using the basis functions which are well-localized in the time-frequency plane as feature extractors.**

I. Summary

The *best-basis* algorithm of Coifman and Wickerhauser [3] was developed mainly for signal compression. This method first expands a given signal into a *dictionary* of orthonormal bases, i.e., a redundant set of wavelet packet bases or local sine/cosine bases having a binary tree structure. The nodes of the tree represent subspaces with different time-frequency localization characteristics. Then a complete basis called a *best basis* which minimizes a certain information cost function (e.g., entropy) is searched in this binary tree using the divide-and-conquer algorithm. This cost function measures the flatness of the energy distribution of the signal so that minimizing this leads to an efficient representation (or coordinate system) for the signal. Because of this cost function, the best-basis algorithm is good for signal compression but is not necessarily good for classification or regression problems.

For classification, we need a measure to evaluate the discrimination power of the nodes (or subspaces) in the tree-structured bases. There are many choices for the discriminant measure $\mathcal{D}$ (see e.g., [1]). For simplicity, let us first consider the two-class case. Let $p = \{p_i\}_{i=1}^n$, $q = \{q_i\}_{i=1}^n$ be two nonnegative sequences with $\sum p_i = \sum q_i = 1$ (which can be viewed as normalized energy distributions of signals belonging to class 1 and class 2 respectively in a coordinate system). One natural choice for $\mathcal{D}$ is *relative entropy*: $D(p,q) \triangleq \sum_{i=1}^n p_i \log(p_i/q_i)$. If a symmetric quantity is preferred, one can use the *J-divergence* between p and q: $J(p,q) \triangleq D(p,q) + D(q,p)$. The measures D and J are both *additive*: for any j, $1 \le j \le n$, $\mathcal{D}(p,q) = \mathcal{D}(\{p_i\}_{i=1}^j, \{q_i\}_{i=1}^j) + \mathcal{D}(\{p_i\}_{i=j+1}^n, \{q_i\}_{i=j+1}^n)$. For measuring discrepancies among L distributions, one may take $\binom{L}{2}$ pairwise combinations of $\mathcal{D}$. The following algorithm selects an orthonormal basis (from the dictionary) which maximizes the discriminant measure on the time-frequency energy distributions of classes. We call this a *local discriminant basis* (LDB).

Algorithm 1 *Given L classes of training signals,*

Step 0: *Choose a dictionary of orthonormal bases (i.e., specify QMFs for a wavelet packet dictionary or decide to use either the local cosine dictionary or the local sine dictionary).*

Step 1: *Construct a time-frequency energy map for each class by: normalizing each signal by the total energy of all signals of that class, expanding that signal into the tree-structured subspaces, and accumulating the signal energy in each coordinate.*

Step 2: *At each node, compute the discriminant measure $\mathcal{D}$ among L time-frequency energy maps.*

Step 3: *Prune the binary tree: eliminate children nodes if the sum of their discriminant measures is smaller than or equal to the discriminant measure of their parent node.*

Step 4: *Order the basis functions by their discrimination power and use $k (\ll n)$ most discriminant basis vectors for constructing classifiers.*

For regression problems, we use the same algorithm by modifying Step 2 and 3 above. In Step 2, we compute the prediction (or regression) error at each node instead of the time-frequency energy distributions. In Step 3, we prune the binary tree by comparing the prediction errors of each parent node and the union of its two children nodes: eliminate the children nodes if their prediction error is larger than their parent node. We call the basis so obtained a *local regression basis* (LRB). One disadvantage is that the prediction error is not an additive measure so that the algorithm is slower than the LDB algorithm.

We tested our method using the triangular waveform classification (three-class problem) described in [2]. We first generated 100 training signals and 1000 test signals for each class. Then, we supplied the raw signals to LDA and CART and obtained the misclassification rates 20.90%, 29.87%, respectively, using the test signals. Finally, we computed the LDB from the wavelet packet dictionary with the 6-tap coiflet filter, and supplied five most discriminant coordinates to LDA and CART. The misclassification rates become 15.90% and 21.37%. Note that the Bayes error of this example is about 14% [2]. The details as well as other examples and applications of LDB/LRB can be found in [4], [5], and [6].

References

[1] M. Basseville, *Distance measures for signal processing and pattern recognition*, Signal Processing **18** (1989), no. 4, 349–369.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Chapman & Hall, Inc., New York, 1993, previously published by Wadsworth & Brooks/Cole in 1984.

[3] R. R. Coifman and M. V. Wickerhauser, *Entropy-based algorithms for best basis selection*, IEEE Trans. Inform. Theory **38** (1992), no. 2, 713–719.

[4] R. R. Coifman and N. Saito, *Constructions of local orthonormal bases for classification and regression*, Comptes Rendus Acad. Sci. Paris, Série I **319** (1994), no. 2, 191–196.

[5] N. Saito, *Local Feature Extraction and Its Applications Using a Library of Bases*, Ph.D. thesis, Dept. of Mathematics, Yale University, New Haven, CT 06520 USA, 1994.

[6] N. Saito and R. R. Coifman, *Local discriminant bases*, Mathematical Imaging: Wavelet Applications in Signal and Image Processing (A. F. Laine and M. A. Unser, eds.), Jul. 1994, Proc. SPIE 2303.

The Role of Approximation and Smoothness Spaces in Compression and Noise Removal

Ronald A. DeVore and Vladimir Temlyakov[1]
Dept. Math., Univ. of South Carolina, Columbia, SC 29208, USA

Abstract — **A brief discussion is given of the role of approximation and smoothness spaces in algorithms for noise removal and compression.**

I. Introduction

Compression and noise removal can be viewed as problems of approximation. Because of space limitations, we limit our discussion to cases where approximation takes place in a Hilbert space $\mathcal{H}$ although the theory applies in far greater generality. Let $\{\phi_\lambda\}_{\lambda\in\Lambda}$ be a a complete orthonormal system for $\mathcal{H}$.

II. Linear and Nonlinear approximation

In linear approximation, we approximate by the elements of the linear spaces $X_n := \text{span}\{\phi_\lambda\}_{\lambda\in\Lambda_n}$, $\Lambda_n \subset \Lambda_{n+1} \subset \Lambda$, $n = 1, 2, \ldots$. If $f = \sum_{\lambda\in\Lambda} c_\lambda\phi_\lambda$ then $\sum_{\lambda\in\Lambda_n} c_\lambda\phi_\lambda$ is its best approximation from X_n and $E_n(f) := (\sum_{\lambda\notin\Lambda_n} |c_\lambda|^2)^{1/2}$ the approximation error.

In nonlinear approximation, we fix a number $n \geq 0$ and approximate f by $\sum_{\lambda\in\Lambda_0} c_\lambda\phi_\lambda$, where Λ_0 is an arbitrary subset of Λ with 2^n elements. The best nonlinear approximation is obtained by taking Λ_0 as the set of the 2^n indicies λ for which $|c_\lambda|$ is largest. We denote the nonlinear approximation error by $\sigma_n(f)$.

III. Approximation spaces

What elements $f \in \mathcal{H}$ can be approximated well by these methods. For example, what elements have an approximation error like $O(2^{-n\alpha})$. For $\alpha > 0$, $0 < q \leq \infty$, let $\mathcal{A}_q^\alpha(\text{L})$ denote the set of $f \in \mathcal{H}$ such that $\sum_{n\geq 1}[2^{n\alpha}E_n(f)]^q$ is finite with the usual change to a sup when $q = \infty$. We replace $E_n(f)$ by $\sigma_n(f)$ to get the space $\mathcal{A}_q^\alpha(\text{N})$. Then, f is in $\mathcal{A}_q^\alpha(\text{L})$ if and only if $\sum_{n\geq 1}[2^{n\alpha}(\sum_{\lambda\in\Lambda_{n+1}\setminus\Lambda_n} |c_\lambda|^2)^{1/2}]^q$ is finite. We can characterize $\mathcal{A}_q^\alpha(\text{N})$ only for special q, namely, $q = (\alpha+1/2)^{-1}$ in which case f is in this space if and only if $\sum_{\lambda\in\Lambda} |c_\lambda|^q$ is finite.

IV. Examples

Let $\mathcal{H} = L_2(\mathbb{T}^d)$ be the space of 2π-periodic functions on the torus and $\phi_k := e_k$, $k \in \mathbb{Z}^d$, with $e_k(x) := e^{ik\cdot x}$, $k \in \mathbb{Z}^d$, the complex exponentials. We take $\Lambda_n := \{k : |k| \leq 2^n\}$. Then, the linear approximation problem corresponds to approximation by the partial sums of the Fourier series of f and $f \in \mathcal{A}_q^\alpha(\text{L})$ if and only if f is in the Besov space $B_q^\alpha(L_2(\mathbb{T}^d))$ (when $q = 2$, $\alpha = r$ is an integer, this is equivalent to f in the Sobolev space $W^r(L_2(\mathbb{T}^d))$. For nonlinear approximation by complex exponentials, $f \in \mathcal{A}_q^\alpha(\text{N})$, $\alpha > 0$, $q = (\alpha + 1/2)^{-1}$ if and only if $\sum_{k\in\mathbb{Z}^d} |\hat{f}(k)|^q$ converges; e.g., if $\alpha = 1/2$, the Fourier series of f should converge absolutely (Stechkin's criteria).

Another important example is when $\Lambda_n := \{k \in \mathbb{Z}^d : |k_1 \cdots k_d| \leq n\}$ is the hyperbolic cross. In this case, $\mathcal{A}_q^\alpha(\text{L})$ is a Besov like space with the usual modulus of smoothness replaced by a mixed modulus [1].

[1]This work was supported by ONR Contract N0014-91-J1343.

V. Wavelets Examples

Let $\mathcal{H} = L_2(\mathbb{R})$ and $\phi \in \mathcal{H}$ be a univariate scaling function with orthonormal shifts $\phi(\cdot - j)$, $j \in \mathbb{Z}$, which generates the orthogonal wavelet ψ. The functions$\psi_{j,k} := 2^{k/2}\psi(2^k \cdot -j)$, $j, k \in \mathbb{Z}$ are a complete orthonormal system for $\mathcal{H}$. For $\Lambda_n := \{(j,k) : j \in \mathbb{Z}, k \leq n\}$, the $\mathcal{A}_q^\alpha(\text{L})$ are again Besov spaces $B_q^\alpha(L_2(\mathbb{R}))$ for a range of α depending on ψ. The nonlinear approximation spaces $\mathcal{A}_q^\alpha(\text{N})$ are the Besov spaces $B_q^\alpha(L_q)$ provided $q = (\alpha + 1/2)^{-1/2}$ [2].

There are various multivariate orthonormal basis for $L_2(\mathbb{R}^d)$ which can be constructed from ϕ and ψ. For example, if $d = 2$, the usual orthogonal basis used in wavelet applications consists of the functions $\eta_{j_1,k}(x)\tilde{\eta}_{j_2,k}(y)$, $j_1, j_2, k \in \mathbb{Z}$, with $\eta, \tilde{\eta}$ either ϕ or ψ but not both ϕ. The approximation classes for linear approximation by partial sums of wavelet series with respect to this basis are Besov spaces $\mathcal{A}_q^\alpha(\text{L}) = B_q^\alpha(L_2(\mathbb{R}^2))$. For nonlinear approximation $\mathcal{A}_q^\alpha(\text{N}) = B_q^\alpha(L_q(\mathbb{R}^2))$, $q = (\alpha/2 + 1/2)^{-1}$.

Another wavelet basis, useful in some applications, is given by the tensor products $\psi_{j_1,k_1}(x)\psi_{j_2,k_2}(y)$, $j_1, j_2, k_1, k_2 \in \mathbb{Z}$. Linear approximation here is analogous to hyperbolic cross Fourier approximation [3].

VI. K-functionals

If $Y \subset X$ are two Banach spaces and $f \in X$, then $K(f, t, X, Y) := \inf_{f=b+g} \|b\|_X + t\|g\|_Y$, $t \geq 0$ is called the K-functional of f. The K-functionals for many classical pairs of spaces are characterized. K-functionals can be used to design (optimal) compression and noise removal algorithms [2]. For example, if $X = L_2(\Omega)$ and $Y = W^r(L_2(\Omega))$ with $\Omega \subset \mathbb{R}^d$ a cube or all of $\mathbb{R}^d$, then the best choice g for fixed t is given by linear approximation (for example using the first wavelet basis in Sect. V). If $Y = \mathcal{W}^r(L_2(\Omega)$ is the space of functions with mixed r-th derivative, a best g is given by linear hyperbolic (or tensor product) approximation. If $Y = B_q^\alpha(L_q(\Omega))$, $q = (\alpha/d + 1/2)^{-1}$, then g is given by nonlinear wavelet approximation. These choices of g give linear or nonlinear compression algorithms optimal for corresponding function classes [2].

In noise removal, one uses the K-functional for noisy f; each t gives a noise removal algorithm. Minimizing the expected error with respect to t leads to linear or nonlinear noise removal algorithms such as wavelet shrinkage [2,4].

References

[1] R. DeVore, P. Petrushev, and V. Temlyakov, " Trigonometric approximation with frequencies from the hyperbolic cross," *Mat. Zametkii* , vol. 56, p. 36–63 , 1994.

[2] R. DeVore and B. Lucier, " Fast wavelet techniques for near optimal image processing," *Proc IEEE Mil. Commun. Conf. Tech. J.*, Oct. 1992, IEEE, N.Y..

[3] R. DeVore, S. Konjagin, P. Petrushev, and V. Temlyakov, "Hyperbolic wavelets," preprint.

[4] D.L. Donoho and I.M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, to appear.

Adaptive Signal Representations: How much is too much?

David L. Donoho[1]
[1]University of California, Berkeley
[2]Stanford University (on leave)

Abstract — **Recently, adaptive signal representations in overcomplete libraries of waveforms have been very popular [1, 5]. One naturally expects that in searching through a large number of signal representations for noisy data, one is at risk of identifying apparent structure in the data which turns out to be spurious, noise-induced artifacts. We show how to use penalties based on the logarithm of library complexity to temper the search, preventing such spurious structure, and giving near-ideal behavior.**

I. Adaptive Signal Representations

Over the last five years or so, there has been an explosion of awareness of alternatives to traditional signal representations. Instead of just representing objects as superpositions of sinusoids (the traditional Fourier representation) we now have available alternate dictionaries – signal representation schemes – of which the Fourier dictionary is only the most well-known. Wavelet dictionaries, Gabor dictionaries, Multi-scale Gabor Dictionaries, Wavelet Packets, Cosine Packets, Chirplets, and a wide range of other representations are now available. Each such dictionary $\mathcal{D}$ is a collection of waveforms $(\phi_\gamma)_{\gamma\in\Gamma}$, and we envision a decomposition of a signal s as

$$s = \sum_{\gamma\in\Gamma} \alpha_\gamma \phi_\gamma. \qquad (1)$$

Depending on the dictionary, such a decomposition is a decomposition into pure tones (Fourier dictionary), bumps (wavelet dictionary), chirps (chirplet dictionary), etc.

A key point. The dictionaries we are interested in are all *overcomplete*. The decomposition (1) is then nonunique, because some elements in the dictionary have representations in terms of other elements. This gives us the possibility of adaptation, i.e. of choosing among many representations one which is most suited to our purposes.

II. Best Ortho Basis

Coifman and Meyer have invented some time-frequency dictionaries, wavelet packets and cosine packets, which have a very special structure. Certain structured subcollections of the elements amount to orthogonal bases; one gets in this way a wide range of orthonormal bases (in fact $\gg 2^n$ such orthogonal bases for signals of length n). Coifman and Wickerhauser [1] have proposed a method of adaptively picking from among these many bases, a single orthogonal basis which is the best one. If $(s[\mathcal{B}]_I)$ denotes the vector of coefficients of s in orthogonal basis $\mathcal{B}$, and if we define the "entropy" $\mathcal{E}(s[\mathcal{B}]) = \sum_I e(s[\mathcal{B}]_I)$, where $e(s)$ is a scalar function of a scalar argument, they give a fast algorithm for solving

$$\min\{\mathcal{E}(s[\mathcal{B}]): \quad \mathcal{B} \text{ ortho basis} \subset \mathcal{D}\}$$

The algorithm is fast – it delivers a basis in order $n\log(n)$ time – and in some cases delivers near-optimal sparsity respresentations.

III. Choice of Entropy for De-Noising

Suppose we have observations $y_i = s_i + z_i$, $i = 1, ..., n$, where (s_i) is signal and (z_i) is i.i.d. Gaussian white noise. Suppose we have available a library $\mathcal{L}$ of orthogonal bases, such as the Wavelet Packet bases or the Cosine Packet bases of Coifman and Meyer. We wish to select, adaptively based on the noisy data (y_i), a basis in which best to recover the signal ("de-noising"). Let M_n be the total number of distinct vectors occcuring among all bases in the library and let $t_n = \sqrt{2\log(M_n)}$. (For wavelet packets, $M_n = n\log_2(n)$.)

Let $y[\mathcal{B}]$ denote the original data y transformed into the Basis $\mathcal{B}$. Choose $\lambda > 8$ and set $\Lambda_n = (\lambda\cdot(1+t_n))^2$. Define the entropy functional

$$\mathcal{E}_\lambda(y,\mathcal{B}) = \sum_i \min(y_i^2[\mathcal{B}], \Lambda_n^2).$$

Let $\hat{\mathcal{B}}$ be the best orthogonal basis according to this entropy:

$$\hat{\mathcal{B}} = \arg\min_{\mathcal{B}\in\mathcal{L}} \mathcal{E}_\lambda(y,\mathcal{B}).$$

Define the hard-threshold nonlinearity $\eta_t(y) = y1_{\{|y|>t\}}$. In the empirical best basis, apply hard-thresholding with threshold $t = \sqrt{\Lambda_n}$:

$$\hat{s}_i^*[\hat{\mathcal{B}}] = \eta_{\sqrt{\Lambda_n}}(y_i[\hat{\mathcal{B}}]).$$

Theorem: With probability exceeding $\pi_n = 1 - e/M_n$,

$$\|\hat{s}^* - s\|_2^2 \le (1-8/\lambda)^{-1}\cdot\Lambda_n\cdot\min_{\mathcal{B}\in\mathcal{L}} E\|\hat{s}_{\mathcal{B}} - s\|_2^2.$$

Here the minimum is over all ideal procedures working in all bases of the library, i.e. in basis $\mathcal{B}$, $\hat{s}_{\mathcal{B}}$ *is just* $y_i[\mathcal{B}]1_{\{|s_i[\mathcal{B}]|>1\}}$.

In short, the basis-adaptive estimator achieves a loss within a logarithmic factor of the ideal risk which would be achievable if one had available an oracle which would supply perfect information about the ideal basis in which to de-noise, and also about which coordinates were large or small.

Acknowledgements

I would like to thank Iain Johnstone for permission to discuss and display our joint results.

References

[1] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best-basis selection", *IEEE Trans. Info. Theory* **38** (1992) p. 713-718.

[2] S. Chen and D. Donoho. Basis Pursuit. Technical Report, Department of Statistics, Stanford University.

[3] D.L. Donoho and I.M. Johnstone, *Ideal Time-Frequency Denoising*. Technical Report, Department of Statistics, Stanford University.

[4] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, (1992).

[5] S. Mallat and Z. Zhang. Matching Pursuit with Time Frequency Dictionaries. *IEEE Trans. Sig. Proc.* (1993).

[1]This work was supported by NSF-DMS-92-09130, and by the NASA Astrophysics Data Program.

Tracking long-range dependencies with wavelets

Patrick Flandrin and Patrice Abry[1]

Laboratoire de Physique (URA 1325 CNRS), *ENS Lyon*, 46 allée d'Italie, 69364 Lyon Cedex 07, France

Abstract — **Long-range dependent processes exhibit features, such as $1/f$ spectra, for which wavelets offer versatile tools and provide a unifying framework. This efficiency is demonstrated on both continuous processes, point processes and filtered point processes.**

I. Introduction

Many signals, in many different domains (solid-state physics, biology, turbulence, communications, ...), exhibit $1/f$ spectra which reveal some long-range dependence (LRD). Although considering LRD processes is therefore necessary, this remains a challenging problem (from the point of view of both modeling and analysis), thus calling for new approaches capable of supplementing some of the specific tools developed so far [3].

II. Fractional Brownian motion

Fractional Brownian motion (fBm) is the first well-known example of a continuous and LRD process for which wavelets proved efficient [4, 5, 8, 9, 10]. The main reasons are as follows: 1. although fBm is nonstationary, its wavelet transform is stationary at any scale (this is due in fact to the stationarity of its increments); 2. the Hurst exponent H of a fBm can be deduced from the variance law of details across scales; 3. whereas fBm is LRD, details of a dyadic decomposition are almost uncorrelated. The effectiveness of using wavelets for fBm analysis can be further evidenced by a comparison with more classical techniques devoted to continuous LRD processes. Given a fBm $B_H(t)$ for which $\mathrm{var}B_H(t)$ behaves as $|t|^{2H}$, it is known that the estimation of H requires the use of a refined variance estimator, referred to as the *Allan variance*. It turns out that such an approach amounts to using a Haar wavelet decomposition, with limitations due to the low regularity of the basis functions [4]. While retaining the same principle, more regular wavelets offer therefore a way of generalizing the Allan variance, with an increased performance.

III. Fractal point process - Fractal shot noise

Beyond fBm, wavelets are also efficient for tracking LRDs in point processes. Let us consider $P(t) = \sum_{k=-\infty}^{+\infty} g(t - t_k)$, where the t_k are Poisson distributed, with an intensity $\lambda(t)$. (The usual Poisson process simply corresponds to $g = \delta$ and $\partial\lambda/\partial t = 0$.) A LRD process (referred to as *fractal point process* (FPP) [6]) can be constructed within this model by choosing $\lambda(t)$ to be *fractional Gaussian noise* (i.e., "derivative" of a fBm). Starting from the remark that, for the counting process $N(T)$ associated to a Poisson process, we have always $\mathrm{var}N(T) = \mathbf{E}N(T)$, a departure from Poisson can be revealed by means of the *Fano factor* $F(T) \equiv \mathrm{var}N(T)/\mathbf{E}N(T)$. In the case of a FPP, $F(T) \sim 1 + C.T^{2H-1}$ when T is large and $H > 1/2$ [6]. In analogy with the definition of $F(T)$, a wavelet based Fano factor $WF(j)$ can then be defined [2], as a function of scale j, by using both the variance of the details $d_P[j,n]$ and the average of the approximations $a_P[j,n]$. The result is

$$WF(j) \equiv (2^j)^{\frac{1}{2}}\mathbf{E}d_P^2[j,n]/\mathbf{E}a_P[j,n] \sim 1 + (2^j)^{2H-1},$$

when j is large and the degree of cancellation R is such that $R > H - \frac{1}{2}$. When using the Haar wavelet in WF and the Allan variance in the estimation of F, we have exactly $WF(j) = F(2^j)$. Wavelets offer therefore a way of generalizing the concept of Fano factor and increasing its efficiency when R is larger than 0. Moreover, the proposed generalization allows to deal directly with filtered point processes, what the Fano factor does not. If we consider for instance the model of *fractal shot noise* (FSN) [7], for which $\lambda(t)$ is a constant but $g(t) = t^{-\beta}$ if $0 < A \leq t < B < +\infty$ and 0 elsewhere, we obtain that $WF(j)$ behaves as $2^{-j\beta}$ when $1/A \gg 2^{-j} \gg 1/B$ (with the only condition $R > \beta - \frac{1}{2}$) [2].

IV. Spectral analysis of $1/f$ processes

In any of the above cases, the basic ingredient in the analysis is the variance of the details, which is time-invariant. This leads to a unified perspective in the frequency domain since such a variance reads

$$\mathbf{E}d_x^2[j,n] = 2^j \int_{-\infty}^{+\infty} |\Psi(2^j f)|^2 \, S_x(f)\, df,$$

where $S_x(f)$ is the (average) power spectrum of the analyzed process and $\Psi(f)$ the Fourier transform of the analyzing wavelet. A consequence of this relation is that wavelet analysis is structurally matched to $1/f^\alpha$ spectra and provides an efficient and unbiased estimation of α, as detailed in [1].

References

[1] P. Abry, P. Gonçalvès and P. Flandrin, "Wavelet-based spectral analysis of $1/f$ processes," IEEE–ICASSP–93, pp. III.237–III.240, Minneapolis (MN), 1993.

[2] P. Abry and P. Flandrin, "Wavelet-based Fano factor for long-range dependent point processes," IEEE–EMBS–94, Baltimore (MD), 1994.

[3] J. Beran, "Statistical methods for data with long-range dependence," *Statistical Science*, Vol. 7, No. 4, pp. 404–427, 1992.

[4] P. Flandrin, "Wavelet analysis and synthesis of fractional Brownian motion," *IEEE Trans. on Info. Theory*, Vol. IT-38, No. 2, pp. 910–917, 1992.

[5] P. Flandrin, "Time-scale analyses and self-similar stochastic processes," in *Wavelets and Their Applications* (J. Byrnes and M. Byrnes, *eds.*), Kluwer, to appear.

[6] D.H. Johnson and A.R. Kumar, "Modeling and analyzing fractal point processes," IEEE–ICASSP–90, pp. 1353–1356, Albuquerque (NM), 1990.

[7] S.B. Lowen and M.C. Teich, "Power-law shot noise," *IEEE Trans. on Info. Theory*, Vol. IT-36, No. 6, pp. 1302–1318, 1990.

[8] E. Masry, "The wavelet transform of stochastic processes with stationary increments and its application to fractional Brownian motion," *IEEE Trans. on Info. Theory*, Vol. IT-39, No. 1, pp. 260–264, 1993.

[9] A.H. Tewfik and M. Kim, "Correlation structure of the discrete wavelet coefficients of fractional Brownian motion," *IEEE Trans. on Info. Theory*, Vol. 38, No. 2, pp. 904–909, 1992.

[10] G.W. Wornell, "Wavelet-based representations for the $1/f$ family of fractal processes," *Proc. IEEE*, Vol. 81, pp. 1428–1450, 1993.

[1] flandrin@physique.ens-lyon.fr, pabry@physique.ens-lyon.fr

Wavelet Vector Quantization with Matching Pursuit

G. Davis and S. Mallat[1]

Courant Institute of Mathematical Sciences, 251 Mercer Street, NY NY 10012

I. Introduction

To compute the optimal expansion of signals in redundant dictionary of waveforms is an NP complete problem. We introduce a greedy algorithm, called matching pursuit, that performs a sub-optimal expansion. This algorithm can be interpreted as a shape-gain multistage vector quantization. The waveforms are chosen iteratively in order to best match the signal structures. Matching pursuits are general procedures to compute adaptive signal representations. Applications to speech and image processing with dictionaries of Gabor functions will be shown, in particular for the removal of noises.

II. Matching Pursuit

Let $\mathbf{H}$ be a signal space. We define a dictionary as a redundant family $\mathcal{D} = (g_\gamma)_{\gamma \in \mathbf{\Gamma}}$ of vectors in $\mathbf{H}$, such that $\|g_\gamma\| = 1$. We impose that linear expansion of vectors in $\mathcal{D}$ are dense in $\mathbf{H}$. An example of dictionary is constructed by dilating, translating and modulating a single window function $g(t)$ of unit norm. For any scale $s > 0$, frequency modulation ξ and translation u, we denote $\gamma = (s, u, \xi)$ and define

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g(\frac{t-u}{s}) e^{i\xi t}. \tag{1}$$

The index γ is an element of the set $\mathbf{\Gamma} = \mathbf{R}^+ \times \mathbf{R^2}$. The factor $\frac{1}{\sqrt{s}}$ normalizes to 1 the norm of $g_\gamma(t)$. If $g(t)$ is even, which is generally the case, $g_\gamma(t)$ is centered at the abscissa u. Its energy is mostly concentrated in a neighborhood of u, whose size is proportional to s.

A signal $f \in \mathbf{H}$ does not have a unique representation as a sum of elements of a redundant dictionary. A matching pursuit decomposes f over a set of vectors selected from $\mathcal{D}$, by successive approximations. Let $g_{\gamma_0} \in \mathcal{D}$, we decompose

$$f = < f, g_{\gamma_0} > g_{\gamma_0} + Rf, \tag{2}$$

where Rf is the residual vector after approximating f in the direction of g_{γ_0}. Clearly g_{γ_0} is orthogonal to Rf, hence

$$\|f\|^2 = |< f, g_{\gamma_0} >|^2 + \|Rf\|^2. \tag{3}$$

To minimize $\|Rf\|$, we must choose $g_{\gamma_0} \in \mathcal{D}$ such that $| < f, g_{\gamma_0} > |$ is maximum.

Let us explain by induction, how the matching pursuit is carried further. Let $R^0 f = f$. We suppose that we have computed the n^{th} order residue $R^n f$, for $n \geq 0$. We choose, with the choice function C, an element $g_{\gamma_n} \in \mathcal{D}$ which closely matches the residue $R^n f$

$$|< R^n f, g_{\gamma_n} >| = \sup_{\gamma \in \mathbf{\Gamma}} |< R^n f, g_\gamma >|. \tag{4}$$

The residue $R^n f$ is sub-decomposed into

$$R^n f = < R^n f, g_{\gamma_n} > g_{\gamma_n} + R^{n+1} f, \tag{5}$$

which defines the residue at the order n+1. Since $R^{n+1} f$ is orthogonal to g_{γ_n}

$$\|R^n f\|^2 = |< R^n f, g_{\gamma_n} >|^2 + \|R^{n+1} f\|^2. \tag{6}$$

If we carry this decomposition up to the order m, we obtain

$$f = \sum_{n=0}^{m-1} < R^n f, g_{\gamma_n} > g_{\gamma_n} + R^m f. \tag{7}$$

and the energy conservation

$$\|f\|^2 = \sum_{n=0}^{m-1} |< R^n f, g_{\gamma_n} >|^2 + \|R^m f\|^2. \tag{8}$$

One can also prove [1] that

$$\lim_{m \to +\infty} \|R^m f\| = 0. \tag{9}$$

This iterative procedure can be interpreted as a shape-gain vector quantization in a very high dimensional space and is also equivalent to a projection pursuit [2], used in statistics.

A matching pursuit can be calculated with a fast algorithm [1] that is described in the talk. In the case of a time-frequency dictionary of Gabor function, the signal is decomposed as a sum of time-frequency elements whose scale, position and frequency match the time-frequency structures of the signal. Applications to noise removal have been developed [1] and we are currently using this representation for music analysis. For image processing, we have constructed a dictionary of two-dimensional Gabor waveforms with an orientation selectivity. Decomposition of images and application to noise removal will be demonstrated.

III. References

1. S. Mallat and Z. Zhang, "Matching Pursuit with Time-Frequency Dictionaries", *IEEE Trans. on Signal Processing*, Dec. 1993.

2. P. J. Huber, "Projection Pursuit", *The Annals of Statistics*, vol. 13, No. 2, p. 435-475, 1985.

[1]This work was supported in part by the AFOSR grant F49620-93-1-0102, ONR grant N00014-91-J-1967 and the Alfred Sloan Foundantion

Multiresolution Models For Random Fields and Their Use in Statistical Image Processing

H. Krim, A. S. Willsky and W.C. Karl[1]

Stochastic Systems Group M.I.T. Room 35-437, M.I.T., Cambridge, MA 02139

I. Introduction

In this paper we describe a probabilistic framework for optimal multiresolution processing and analysis of spatial phenomena. Our developed Multiresolution (MR) models are useful in describing random processes and fields. The scale recursive nature of the resulting models, leads to extremely efficient algorithms for optimal estimation and likelihood calculation. These models, described below, have also provided a framework for data fusion, and produced new solutions to problems in computer vision (optical flow estimation), remote sensing (oceanography where dimensional complexity is in thousands), and various inverse problems of mathematical physics.

II. Recursive MR Models

The stochastic models that form the focus for our work are defined on a tree T, where we use the index t to denote a general node of the tree. In our context the nodes of the tree are organized into levels or resolutions, corresponding to different resolutions of representation for the phenomenon of interest. In particular, we can think of the nodes on the tree as 2-tuples, $(m(t), n(t))$, where $m(t)$ denotes the scale of the node t and $n(t)$ the spatial location corresponding to that node. In describing images or 2-D signals, $m(t)$ and $n(t)$ may be vectors themselves, describing scales and translational locations in the two coordinate directions.

The models of interest in our work are scale-recursive Markov models on T. Specifically, let 0 denote the root node of the tree (i.e., the single node at the coarsest scale), let $t\gamma$ denote the parent of node t, and let $t\alpha_1, \cdots, t\alpha_p$ denote the descendants of t (where in general the number of descendants may vary from node to node). Then the model is given by

$$\mathrm{x}(t) = A(t)\mathrm{x}(t\gamma) + B(t)w(t),$$

where $w(t)$ is a zero-mean, unit variance, white process on T which is independent of $\mathrm{x}(0)$ and $A(t)$ and $B(t)$ are matrices that may (and frequently do) vary with t or, at least, $m(t)$. For example, the modeling of process with particular scaling laws, such as fractals, typically involve the use of noise gains that decrease geometrically with scale.

Defined in this way, $\mathrm{x}(t)$ is obviously a Markov random field on T, and, moreover, given the value of $\mathrm{x}(t)$, the values of $\mathrm{x}(\cdot)$ on the numerous disjoint subtrees extending from the node t are mutually independent. It is this fact that leads directly to efficient algorithms for multiresolution signal and image analysis. Specifically, consider the following set of multiscale measurements:

$$\mathrm{y}(t) = C(t)\mathrm{x}(t) + v(t),$$

where $v(t)$ is a zero-mean, unit variance white noise process independent of $\mathrm{x}(t)$. Note that this model allows for measurements at multiple resolutions and also allows for nonstationary, sparse and irregular measurements (in which case $C(t)$ certainly varies with t and is zero except at selected nodes at which measurements are available). This scale-recursive model for $\mathrm{x}(t)$ and the associated measurement model for $\mathrm{y}(t)$ admit an extremely efficient algorithm for estimating $\mathrm{x}(t)$ throughout the tree given all of the available data. The fine-to-coarse processing step computes the optimal estimate at each node given the measurement at that node and at nodes in its descendent subtree. Being highly parallelizable and, thus well- matched to hypercube architectures, the algorithm is still extremely efficient even on a serial machine.

Note that the total number of nodes in the tree is a rather small multiple of N ($2N$ for dyadic trees used for 1- D signals and $(4/3)N$ for quadtrees frequently used in image processing) and the total computational complexity of the estimation algorithm is $O(N)$ –i.e., in image processing problems, it has constant per-pixel computational complexity independent of image size while producing estimates at a full set of resolutions.

III. Applications

The importance of these algorithms is further marked by the wealth of physical phenomena and applications whose models are fraught with a computational complexity which could otherwise be prohibitive. These algorithms have been very successfully applied to the problem of "optical flow" estimation from image sequences and where the smoothness penalty corresponded to a prior fractal model. In addition, a performance similar to that of exact MRF likelihood calculation has also resulted even in problems where nonstationary phenomena were present. Our latest applications of these methods involved very high dimensional *Oceanography* problems where the processing efficiency of sparse altimetry data from Topex/Poseidon satellite resulted in maps of of sea level variations along with error statistics.

Finally, an area in which we believe our methods should be particularly well-matched, is that of image reconstruction and inverse problems in which blurred, integrated, or indirect measurements of a random field are to be used in order to estimate the field or to perform other tasks such as texture discrimination, anomaly; detection, etc. In particular we apply the multiresolution modeling methods ealier developed, to the problems of modeling the statistical variability of synthetic aperture radar (SAR) imagery and then using these models for the discrimination of targets from clutter. We demonstrate that statistical fluctuations are well- captured by models of the type that we have described, with significant differences between the models for clutter and for targets, both in the model parameters and in the statistics of the scale-to-scale detail process w(t) (which is Gaussian for targets and log- Rayleigh for clutter).

[1]This was supported in part by the Army Research Office (DAAL-03-92-G-115), Air Force Office of Scientific Research (F49620-92-J-2002) and National Science Foundation (MIP-9015281).

POSTER SESSION I

Neural Network Approximation and Estimation of Functions

Gerald H. L. Cheang[1]
Dept. of Statistics, Yale University, Box 208290, Yale Station, New Haven, CT 06520-8290

Abstract — **Approximation and estimation bounds were obtained by Barron (1992, 1993 and 1994) for function estimation by single hidden-layer neural nets. This paper will highlight the extension of his results to the two hidden-layer case. The bounds derived for the two hidden-layer case depend on the number of nodes T_1 and T_2 in each hidden-layer, and also on the sample size N. It will be seen from our bounds that in some cases, an exponentially large number of nodes, and hence parameters, is not required.**

I. Introduction

A single hidden-layer feedforward sigmoidal network is a family of functions $f_T(x)$ of the form

$$f_T(x,\theta) = \sum_{i=1}^{T} c_i \phi(a_i \cdot x - b_i), x \in R^d$$

parametrized by $\theta = (a_i, b_i, c_i)_{i=1}^M$ with internal weight vectors a_i in R^d, internal location parameter b_i in R, external weights c_i, and ϕ any sigmoidal function with distinct finite limits at $+\infty$ and $-\infty$. Such a network has d inputs, T hidden nodes and a linear output unit. It implements the ridge-function $\phi(a_i \cdot x - b_i)$ on the nodes in the hidden layer. The network model can be used to approximate target functions $f(x)$ defined over bounded subsets of R^d and to estimate the function based on data $(X_i, Y_i)_{i=1}^N$, a random sample from a joint probability distribution $P_{X,Y}$ with $f(x) = E[Y_i|X_i = x]$.

This presentation will be concerned with extensions for approximation and estimation bounds for two hidden-layer sigmoidal networks. Such a network takes the form

$$f_{T_1,T_2}(x,\theta) = \sum_{i=1}^{T_1} c_i \phi(\sum_{j=1}^{T_2} a_{ji} \phi(\omega_{ji} \cdot x + b_{ji}) - d_i), x \in R^d$$

There are T_1 nodes in the outer layer and T_2 nodes in the inner layer, giving a total of $T_1 + T_1 T_2$ nodes. It is parametrized by $\theta = (c_i, d_i, b_{ji}, \omega_{ji}, a_{ji})_{i=1\,j=1}^{T_1\;T_2}$.

II. Approximation Bounds

The approximation bound for the single hidden-layer case was already obtained by Barron (1992 and 1993) for function estimation by single hidden-layer neural nets. This paper will highlight the extensions to the two hidden-layer case. We will show that by using a family of two hidden-layer neural nets to approximate a target function, we are able to approximate some classes of functions that are not known to be approximable by single hidden-layer neural nets. Barron's (1992) L_2 approximation bound was $O(C_{f,hs}/\sqrt{T})$ where T is the number of nodes, and $C_{f,hs}$ is the variation of the target function with respect to the half-spaces. We show that two hidden-layer nets can accurately approximate functions that have bounded variation with respect to larger classes of sets. For example, if the target function f has bounded variation with respect to a class of ellipsoids, then the L_2 approximation error is

$$||f - f_{T_1,T_2}||_2 \leq V/\sqrt{T_1} + K/T_2^{1/6} \qquad (1)$$

where V depends on the variation property of the target function and K depends on the curvature of the ellipsoids, when such a function is approximated by a two layer neural net with T_1 nodes in the outer layer and T_2 nodes in the inner layer. The indicator of a ball is an example of a function that apparantly cannot be approximated accurately by a single hidden-layer net (with a linear output unit) but is approximated well with two layers.

III. Estimation Bounds

In deriving the estimation bound, the target function is assumed to be estimated from the data $(X_i, Y_i)_{i=1}^N$, a random sample of size N from a joint probability distribution $P_{X,Y}$ with $f(x) = E[Y_i|X_i = x]$. Barron's (1994) result for the single hidden-layer case was $O(\frac{C_f^2}{T}) + O(\frac{Td}{N}\log N)$, where d in the dimension of the input, N is the sample size and T is the number of nodes. In our extension to the two hidden-layer case, the overall mean squared estimation error in terms of the best approximation error, the dimension of the parameter space m_{T_1,T_2} and the sample size N is bounded by

$$O(||f - f_{T_1,T_2}||_2^2) + O(m_{T_1,T_2}\log N/N) \qquad (2)$$

In (2), the first term is obtained from (1). It can be seen from our bounds that in some cases, an exponentially large number of nodes, and hence parameters, is not required. Complexity regularization, and a calculation of an index of resolvability, as in Barron (1994), is used in the derivation of our estimation bound.

References

[1] A. R. Barron, "Neural net approximation", *Proc of the 7th Yale workshop on adaptive and learning systems*, 1992

[2] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function", *IEEE Transactions on Information Theory* , vol. 39, pp. 930– 944, 1993

[3] A. R. Barron, "Approximation and estimation bounds for artificial neural networks", *Machine Learning* , vol. 14, pp. 113–143, 1994

[1]email : cheang@stat.yale.edu

Markov Chains and Random Walks in Data Communication Receivers

John W. Craig
Interstate Electronics Corporation, 1001 E. Ball Road, P.O. Box 3117, Anaheim, CA 92803

***Abstract* – For a loss of lock indicator using an up/down counter, the probabilities of true and false declarations of in-lock and out-of-lock are calculated.**

I. Introduction

In many data communication receivers up/down counters are used as a critical part of the processing to determine whether the symbol timing and/or carrier phase tracking phase-locked loops are in-lock or out-of-lock, and it is necessary to calculate the various probabilities for true and false indications of in-lock or out-of-lock. A random walk along a line (which is viewed as a Markov chain) is an exact model of an up/down counter. The random walk has N states, and in this application one end is a partially reflecting barrier, and the other end is an absorbing barrier or sink. Previously published analyses have focused on finding the average time to make a declaration and its variance. In this paper we concentrate on finding the probabilities of making a true or a false declaration within a certain number of symbol intervals or within a certain length of time.

II. Calculation of Probabilities

Two different approaches are required to calculate the desired probabilities. The first is through the transfer function of the equivalent signal flow graph of the random walk [1], and the second is by means of the diagonal form of the tridiagonal state transition matrix [2] that has been found to have distinct eigenvalues. Since the random walk has a finite number of states, the transfer function is, of course, rational. In many published results of this type, the expression for the transfer function has a removable singularity, but here we give explicit, general expressions for the numerator and denominator polynomials (without common roots) of the transfer functions. Since the numerical factors in all the polynomial coefficients are integers, they can be readily and exactly calculated. For a general random walk along a line, the denominator of the transfer function satisfies a second-order difference equation whose solution is the general polynomial mentioned above.

The probabilities of interest are given by the coefficients in the power series expansion of the transfer function about $Z = 0$. The coefficient of the m^{th} power of Z is the probability of being at the chosen position in exactly m steps. It is also given by a certain element in the m^{th} power of the state transition matrix of the Markov chain. Often one is interested in the cumulative probability or the probability of being at a certain position or state in any number of steps less than or equal to m. This is given by the coefficient of the m^{th} power of Z in the power series expansion of the transfer function divided by $1-Z$.

III. Numerical Calculations

Calculation of these probabilities is a difficult numerical problem when the number of states in the random walk is greater than 10 or so and/or the number of steps is in the hundreds. The difficulty is compounded when the number of steps is in the hundreds of thousands or millions, and there are practical situations where this is required. For the number of states up to 100 and the number of steps up to 500, it has been found that the power series expansion capability of Mathematica does an excellent job in calculating the probabilities, which are produced as exact fractions when the state transition probabilities are read in as fractions. For situations requiring hundreds of thousands of steps, the eigenvalue expansion or diagonal form of the state transition matrix has been used with some success to compute powers of this matrix. However, with the double precision subroutines available for making this expansion, the generated orthogonal eigenvector matrix is often so close to being singular that its required inverse cannot be calculated reliably; thus this approach breaks down. At this time it is unknown whether the singular nature is caused by numerical imprecision or whether it is inherent in the problem for some values of state transition probability. However, the former is suspected. Several numerical examples are given.

References

[1] R. W. Sittler, "Systems Analysis of Discrete Markov Processes," *IRE Trans. Circuit Theory*, vol. CT–3, pp. 257–266, 1956.

[2] L. Takacs, *Stochastic Processes*, pp. 5–11, London, Methuen, 1960.

MMSE Parameter Estimation of Exponentially Damped Sinusoids

Hsiang-Tsun Li and Petar M. Djurić

Department of Electrical Engineering, State University of New York at Stony Brook, Stony Brook, NY 11794, USA.

Abstract — **An efficient iterative MMSE algorithm that estimates the parameters of exponentially damped sinusoids embedded in Gaussian noise is proposed.**

I. Introduction

In many engineering and scientific problems the observed measurements are modeled as exponentially damped sinusoids distorted by additive noise. A difficult but interesting problem has always been the estimation of the nonlinear parameters of these signals, the frequencies and the damping factors. There have been a variety of approaches for estimation, most of them revolving around the maximum likelihood (ML) principle. Here we propose a method that yields the minimum mean square estimates (MMSE) of the frequencies and damping factors with all the remaining parameters of the model being considered nuisance.

II. Problem Statement

We assume that an $N \times 1$ data vector $\mathbf{y}$ represents m exponentially damped sinusoids embedded in white Gaussian noise. In particular, $\mathbf{y}$ is given by $\mathbf{y} = \mathbf{H}\mathbf{a} + \mathbf{w}$ where $\mathbf{H}$ is an $N \times 2m$ matrix whose columns span the signal space, $\mathbf{a}$ is a vector of amplitudes, and $\mathbf{w}$ a noise vector with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The matrix $\mathbf{H}$ is defined by

$$\mathbf{H} = [\mathbf{s}_{1c}\ \mathbf{s}_{1s}\ \mathbf{s}_{2c}\ \mathbf{s}_{2s}\ \cdots\ \mathbf{s}_{mc}\ \mathbf{s}_{ms}]$$

$$\mathbf{s}_{kc}^T = [1 \quad e^{-\alpha_k}\cos(2\pi f_k) \quad \cdots \quad e^{-\alpha_k(N-1)}\cos(2\pi f_k(N-1)]$$
$$\mathbf{s}_{ks}^T = [0 \quad e^{-\alpha_k}\sin(2\pi f_k) \quad \cdots \quad e^{-\alpha_k(N-1)}\sin(2\pi f_k(N-1)].$$

All the signal parameters are unknown as is the noise variance σ^2. Given the observations $\mathbf{y}$, the objective is to estimate the nonlinear parameters f_k and α_k, $k = 1, 2, \ldots m$, of the signals.

III. MMSE Estimator

Let the unknown frequencies and damping factors be denoted by $\mathbf{f}$ and α, respectively. The MMSE estimates are given by

$$\hat{\mathbf{f}} = \int_{\mathbf{f},\alpha} \mathbf{f}\, p(\mathbf{f}, \alpha \mid \mathbf{y}) d\alpha d\mathbf{f}, \quad \hat{\alpha} = \int_{\alpha,\mathbf{f}} \alpha\, p(\mathbf{f}, \alpha \mid \mathbf{y}) d\mathbf{f} d\alpha \quad (1)$$

where $p(\mathbf{f}, \alpha \mid \mathbf{y})$ is the a posteriori probability density function of the frequencies and damping factors. Note that the amplitudes and the noise variance σ^2 have been integrated out analytically. The integrals (1) are $2m$-dimensional, and as such, would require reliance on numerical techniques for high dimensional integration. An alternative is to resort to an iterative approach similar in philosophy to the expectation-maximization [1] and alternating projections [2]. To be more specific, let i denote the current iteration, and $\hat{f}_j^{(i)}$, $\hat{\alpha}_j^{(i)}$, the current estimates of f_j and α_j, $j = 1, 2, \cdots, m$. Then, if we approximate the a posteriori density $p(\mathbf{f}, \alpha \mid \mathbf{y})$ by

$$p(\mathbf{f}, \alpha \mid \mathbf{y}) \simeq p(f_k, \alpha_k \mid \mathbf{y}, \hat{\mathbf{f}}_{(-k)}^{(i)}, \hat{\alpha}_{(-k)})^{(i)} \prod_{\substack{j=1 \\ j \neq k}}^{m} \delta(f - \hat{f}_j^{(i)}) \delta(\alpha_j - \hat{\alpha}_j^{(i)})$$

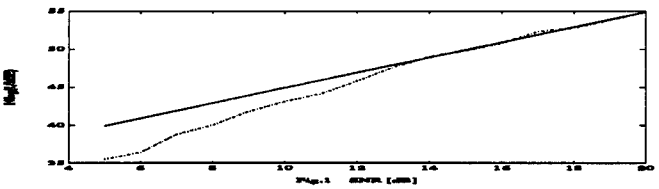
our $2m$-dimensional integrals would reduce to 2-dimensional integrals. $\hat{\mathbf{f}}_{(-k)}^{(i)}$ and $\alpha_{(-k)}^{(i)}$ denote the estimates at the i-th iteration of all the frequencies and damping factors except the ones of the k-th signal. For example, the 2-dimensional integrals for the frequencies have the form

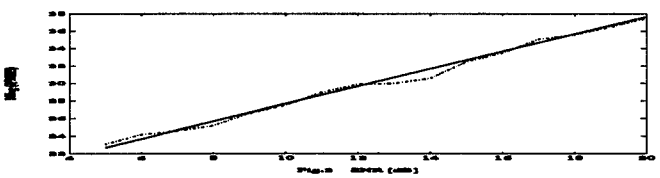
$$\hat{f}_k^{(i)} = \int_{f_k,\alpha_k} f_k\, p(f_k, \alpha_k \mid \mathbf{y}, \hat{\mathbf{f}}_{(-k)}^{(i)}, \alpha_{(-k)}^{(i)}) d\alpha_k df_k. \quad (2)$$

The integrals for the damping factors are similar to (2). The method is based on solving integrals such as (2) until convergence of the estimates is achieved.

IV. Simulation Result

In the computer experiment we generated two damped sinusoids in noise. The amplitude vector was equal to $\mathbf{a}^T = [1\ 0\ 1\ 0]$, the normalized frequencies were $f_1 = 0.16$ and $f_2 = 0.26$, and the damping factors $\alpha_1 = 0.2$ and $\alpha_2 = 0.1$. The SNR was varied between 5 and 20 dB in steps of 1 dB. For each SNR we simulated 100 realizations. The two dimensional integration was carried out by an adaptive importance sampling technique from [3]. The results for f_2 and α_2 are shown in Figures 1 and 2, respectively. Similar results were obtained for f_1 and α_1. In each figure, the solid line represents the Cramer-Rao (CR) bounds, and the other, the mean squared error of our estimates.





This work was supported by the National Science Foundation under Award No. MIP-9110628.

References

[1] M.Feder and E. Weinstein, "Parameter Estimation of Superimposed Signals Using the EM Algorithm", *IEEE Trans. on ASSP*, vol 36, pp. 477-489, Apl. 1988.

[2] I.Ziskind and M. Wax, "Maximum Likelihood Localization of Multiple Sources by Alternating Projection", *IEEE Trans. on ASSP*, vol 36, pp. 1553-1560, Oct. 1988.

[3] D.E. Johnston and P.M. Djurić, "Bayesian Detection ans MMSE Frequency Estimation of Sinusoidal Signals via Adaptive Importance Sampling", *International Symposium on Circuits and Systems*, June 1994.

Adaptive Edge Detection in Compound Gauss-Markov Random Fields Using the Minimum Description Length Principle

Mário A. T. Figueiredo and *José M. N. Leitão*

Instituto de Telecomunicações, and Departamento de Engenharia Electrotécnica e de Computadores.
Instituto Superior Técnico, 1096 Lisboa Codex, PORTUGAL

Abstract – **Edge location in compound Gauss-Markov random fields is formulated as a parameter estimation problem; since the number of parameters is unknown, a minimum-description-length (MDL) criterion is proposed.**

I. Introduction

Compound Gauss-Markov random field (CGMRF) models allow for edge-preserving Bayesian image restoration/reconstruction using continuous (Gaussian) statistical models together with a binary (hidden) edge field [1]. The CGMRF approach to simultaneous edge detection and image restoration involves two random fields: one (intensity field) representing the image to be restored and another one signaling edge elements. To perform joint *maximum a posteriori* (MAP) estimation of both the image and its edges, some prior model has to be specified for the edge process. This prior is usually not explicitly stated; instead, a joint intensity-edge prior is directly considered [1], [2], [3].

Our approach does without the specification of any prior for the line process by adopting a new perspective: we interpret edge locations as (deterministic but unknown) parameters of the original image prior model. Locating edges is then a parameter estimation problem with a salient feature: unknown number of parameters (edges). This fact places the problem in a class to which Rissanen's *minimum description length* (MDL) principle has been successfully applied [4].

We propose an MDL-type edge location criterion for image restoration based on a CGMRF model; it contains no edge-related parameters, such as detection penalty, which appear (and have to be specified) in other types of models.

II. The MAP Estimate and the CGMRF Model

Let $\mathbf{x}$ be a noncausal CGMRF, modelling the original image to be estimated, and $\mathbf{y}$ a linear observation (LO) of $\mathbf{x}$, contaminated by additive white Gaussian noise (AWGN). Let $\mathbf{l}$ be the (hidden) binary edge field (line process); its elements, placed on an interpixel dual grid, indicate whether *bonds* between elements of $\mathbf{x}$ are broken or not. What is usually sought for is the joint MAP estimate of $\mathbf{x}$ and $\mathbf{l}$, given $\mathbf{y}$, which is the mode of $p(\mathbf{x},\mathbf{l}|\mathbf{y})$ or of $p(\mathbf{x},\mathbf{y}|\mathbf{l})\,p(\mathbf{l})$. Here, $p(\mathbf{y},\mathbf{x}|\mathbf{l})$ is the joint PDF of $\mathbf{x}$ and $\mathbf{y}$ given a certain edge configuration, and $p(\mathbf{l})$ the prior of the line process. Notice that $\mathbf{l}$ can be seen as a parameter of $p(\mathbf{x}|\mathbf{l})$ or of $p(\mathbf{x},\mathbf{y}|\mathbf{l})$; under the CGMRF–LO–AWGN assumptions, these are both Gauss proability density functions which depend on $\mathbf{l}$.

III. The Minimum Description Length Principle

The MDL principle generalizes the *maximum likelihood* (ML) criterion to cases where a parameter vector θ of unknown dimension k is to be estimated [4]. The (joint) MDL estimate of k and θ, given observed data $\mathbf{z}$, is

$$(\widehat{k},\widehat{\theta}) = \arg\min_{k,\theta} \{-\log_2 p(\mathbf{z}|\theta,k) + L(\theta|k) + L(k)\}, \qquad (1)$$

where $L(\theta|k)$ and $L(k)$ are, respectively, the code lengths for θ (given that it is k-dimensional) and for k itself; for further details see [4] and the references therein. As usual, $L(k)$ is here considered constant and dropped.

IV. Proposed Approach

To abandon any prior assumption (expressed in $p(\mathbf{l})$) about the edge field, we interpret edge locations as *priorless* parameters of $p(\mathbf{x}|\mathbf{l})$.

Let the locations of all (say k) signaled edges be collected in a k-dimensional parameter vector θ. Writing $p(\mathbf{x}|\mathbf{l})$ is equivalent to writing $p(\mathbf{x}|\theta,k)$ since θ is just a compact code for $\mathbf{l}$. In a first order model [1], [2], [3], and taking $M \times N$ size images, we need $\log_2(MN)$ bits to code each edge location plus 1 bit to distinguish horizontal from vertical edges. Accordingly, $L(\theta|k) = k\log_2(2MN)$.

In the presence of both $\mathbf{x}$ and $\mathbf{y}$, the MDL estimates of k and θ could be obtained by considering $\mathbf{z} = (\mathbf{x},\mathbf{y})$ and inserting $p(\mathbf{x},\mathbf{y}|\theta,k) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}|\theta,k)$ (notice that $p(\mathbf{y}|\mathbf{x},\theta,k) = p(\mathbf{y}|\mathbf{x})$) plus the parameter code length $L(\theta|k)$ into the MDL criterion (1). This yields an MDL estimation criterion for k and θ (*i.e.* the number of edges and their locations):

$$(\widehat{k},\widehat{\theta}) = \arg\min_{k,\theta} \{k\log_2(2MN) - \log_2 p(\mathbf{x},\mathbf{y}|\theta,k)\}. \qquad (2)$$

The criterion specified by (2) has an intrinsic difficulty lying in the fact that $\mathbf{x}$ is not observed (is missing); i.e. it can be classified as MDL parameter estimation from incomplete data. To deal with (2), we have developed a modified version of the *expectation-maximization* (EM) algorithm [5]; further details are presented in [6]. Although it is a suboptimal scheme, the results obtained show the ability of the proposed criterion to adapt to the image edge structure [6].

References

[1] F. Jeng and J. Woods, "Image estimation by stochastic relaxation in the compound Gaussian case", *in Proc. of ICASSP'88*, pp. 1016–1019, New York, 1988.

[2] T. Simchony, R. Chellappa, and Z. Lichtenstein, "Graduated nonconvexity algorithm for image estimation using compound Gauss Markov field models", *in Proc. of ICASSP'89*, pp. 1417–1420, Glasgow, 1989.

[3] J. Zerubia and R. Chellappa, "Mean field annealing using compound Gauss-Markov random fields for edge detection and image estimation", *IEEE Trans. on Neural Net.*, vol. 4, pp. 703–709, July 1993.

[4] J. Rissanen, *Stochastic Complexity in Stastistical Inquiry*, World Scientific, Singapore, 1989.

[5] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.

[6] M. Figueiredo, J. Leitão, "Adaptive Discontinuity Location in Image Restoration", *in Proc. of Intern. Conf. on Image Processing – ICIP'94*, Austin, 1994.

Maximized Mutual Information Using Macrocanonical Probability Distributions

Robert L. Fry

The Johns Hopkins University/Applied Physics Laboratory, Laurel, MD 20723, USA

Abstract – **A maximum entropy formulation leads to a neural network which is factorable in both form and function into individual neurons corresponding to the Hopfield neural model. A maximized mutual information criterion dictates the optimal learning methodology using locally available information.**

I. INTRODUCTION

A biological model is developed here in which neurons computationally realize a multi-dimensional hypothesis testing function implemented on a neural field $Q=\{q_1\ q_2\ \ldots\ q_M\}$ of propositions $\mathbf{y}\epsilon B^M$, $B=\{0,1\}$, $M\epsilon Z$ which become defined through learning. Each neural output y_i describes a conjunctive component of a compound proposition $q=y_1 \cdot y_2 \cdot y_3 \cdot \ldots \cdot y_M$ posed to observed input originating from arbitrary sources. Answers represent decisions which are in turn provided as individual neural output indications (action potentials). Learning within the neural field is realized by the definition of the elemental propositions which in turn correspond to those propositions which serve to maximize the channel capacity between input ensembles and the ensemble of recalled states. The operational objective of the field corresponds to the search for global minima of a quadratic energy function $E_Q=\Sigma e_i$ which parameterizes the *a posteriori* Gibbs distribution as conditioned on the input vector cue. The negative of the respective neuron energies $-e_i$ correspond to the statistical evidence using in determining the probability of generating an action potential.

II. MAXIMUM ENTROPY (ME) FORMULATION

It is assumed that the neural field Q is capable of extracting information from external inputs $\mathbf{x}\epsilon B^N$ and its own outputs $\mathbf{y}$ through a set of sampling functions F. The field Q uses the sampled data to estimate moments on the defined sampling functions which it in turn uses to form the joint ME distribution $P(\mathbf{x},\mathbf{y})$. As such, this probability is a property of the observing ensemble Q as it should be since probability is deemed to be a property of the observer [1]. The computed moments serve to realize $P(\mathbf{x},\mathbf{y})$ as a unique network ME distribution or equivalently a Gibbs distribution parameterized by synaptic connection weights which are in fact the Lagrange multipliers for the ME distribution.

III. MAXIMIZED MUTUAL INFORMATION (MMI)

The Gibbs Mutual Information Theorem as derived in [2] is applied to the composite network distribution which then serves to constrain the network architecture and signal processing required to approximate the MMI criterion between the input ensemble $\mathbf{x}$ and the output ensemble $\mathbf{y}$.

The use of an MMI criterion serves to optimize the storage capacity of the network which is given by the entropy $H(\mathbf{y})$ of the network storage capacity. $H(\mathbf{y})$ has a theoretical maximum of M bits which can asymptotically be obtained using the MMI criterion. This value can change once dynamical considerations are included. Degradation of the storage capacity due to input noise is not considered here, but has been considered elsewhere. Degraded decipherability of the input code by an observer of the output neural code trying to guess the input code can be attributed to either noise, the many-to-one compression imposed by the condition that $N>M$, or both.

IV. RESULTS

Together ME and MMI lead to a neural field which is factorable in both form and function into component computational entities which correspond to the Hopfield neuron model ([3]) including decision threshold, action potential realization, Hebbian learning, sigmoidal transfer characteristic, and conditionalized principal component analysis using a simple modification of an equation originally described by Oja [4].

A diffusion-based search scheme using Langevin's equation in conjunction with the neural field energy E_Q is shown to lead to the FitzHugh-Nagumo neuron activation model. Synchronous activation patterns observed in biological assemblies of neurons can then be described as an asymptotic periodic Markov chain realized through a Gibbs-sampler computational paradigm. Quantitative details of this process are alluded to.

REFERENCES

[1] R. T. Cox, *The algebra of probable inference*, The Johns Hopkins Press, Baltimore, 1961.

[2] R. L. Fry, "Observer-participant models of neural processing," *IEEE Trans. Neural Networks*, In Press.

[3] R. L. Fry, "Neural processing of information," paper presentation at the 1994 Int. Symp. on Info. Theory, Trondheim, Norway.

[4] E. Oja, "A simplified neuron model as a principal component analyzer," *J. of Math. Biol.* **15**, 267-273., 1982

Sample Path Description of Gauss Markov Random Fields [1]

Sauraj Goswami and José M. F. Moura
Electrical Eng. Dept., Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract — **We provide a characterization of Gauss Markov random fields in terms of partial differential equations with random forcing term. Our method consists of obtaining a concrete representation of an abstract stochastic partial differential equation using some results from the theory of vector measures.**

I. Preliminary Background

To fix notations, let $Y_u, u \in K$ (K compact subset of R^n) be a random field and let D be an open subset of K. Let Γ be the boundary of D. Then it is well known that the field Y_u is Markov with respect to D if

$$\mathrm{E}[Y_u Y_v \mid \sigma(\Gamma)] = \mathrm{E}[Y_u \mid \sigma(\Gamma)]\mathrm{E}[Y_v \mid \sigma(\Gamma)] \tag{1}$$

where $u \in D$ and $v \in D^c$ and $\sigma(\Gamma)$ is the usual germ-field given by

$$\sigma(\Gamma) = \cap\{\sigma(O): O \text{ open and } O \supset \Gamma\} \tag{2}$$

Thus D and D^c are conditionally independent given knowledge of the boundary.

For Gaussian fields , conditioning on $\sigma(A)$ ($A \subset K$) is projection onto the closed subspace generated by $Y_u, u \in A$, instead of the larger subspace of all L^2 functions measurable with respect to $\sigma(A)$ (see [2]). Hence, the Markov property can be formulated in terms of projection on these smaller Hilbert spaces. We introduce the spaces

$$H(K) = \text{closed subspace generated by } Y_u, u \in K \tag{3}$$

and the corresponding reproducing kernel Hilbert space $\mathcal{H}(K)$. It is well known that $H(K)$ and $\mathcal{H}(K)$ are isometrically isomorphic through the mapping, $J: H(K) \mapsto \mathcal{H}(K)$

$$JY(t) = \mathrm{E}Y_t \; t \in K. \tag{4}$$

II. Sample Path Characterization

We assume that $C_0^\infty(K)$ is dense in $\mathcal{H}(K)$. For $u, v \in C_0^\infty(K)$, we can write the inner product of $\mathcal{H}(K)$ in the form

$$< u, v >_{\mathcal{H}(K)} = (Pu, v)_{L^2} \tag{5}$$

where P is a differential operator written in the divergence form (see [3]).

In order to derive a sample path characterization we use the well known technique for associating a generalized random field ξ to our ordinary random field Y through the following formula

$$\xi(\phi) = \int_K Y(u, \omega)\phi(u)\, du. \tag{6}$$

A generalized random field can be regarded as a linear operator from C_0^∞ to a space of L^2 random variables. With every generalized field, there is an associated dual field. The dual field ξ^* is also a linear operator from C_0^∞ to L^2 random variables such that

$$\mathrm{E}[\xi(u)\xi^*(v)] = \int_K u(t)v(t)\, dt \tag{7}$$

Kallianpur and Mandrekar [5] has shown that an ordinary random field is Markov if and only if the associated generalized random field is Markov. This enables us to study the generalized random field and then transfer back its properties to the associated ordinary random field.

Now, the generalized field is Markov if the dual field ξ^* is local, (see [1]) in the sense that if $\mathrm{supp}u \cap \mathrm{supp}v = \phi$, then $\mathrm{E}(\xi^*(u)\xi^*(v)) = 0$. Locality of the dual field implies that $\mathrm{E}[\xi^*(u)\xi^*(v)] = (Pu, v)_{L^2}$ where P is the same differential operator associated with inner product of the RKHS of the ordinary random field (see equation (5)).

We know [1] that ξ satisfies the following abstract equation

$$\xi(Pu) = \xi^*(u). \tag{8}$$

We show that the dual of the generalized field is intimately related to J^{-1} (J is defined in equation (4)). Under some integrability condition we further show that the mapping J^{-1} is weakly compact. A weakly compact mapping from $L^1(\mu)$ to a Hilbert space is Riesz representable (see [4]). Therefore, we can write equation (8) in the following weak form

$$\int Y(t, \omega)Pu(t)\, dt = \int \epsilon(t, \omega)u(t)\, dt \tag{9}$$

where $u(t) \in C_0^\infty(K)$.

When the support of u is in D, a subset of K, we relate $\xi^*(u)$ to minimum mean square error. In particular, we show that $\xi^*(u)$ lies in the closed subspace generated by $(Y_u - \mathrm{E}[Y_u \mid \sigma(\Gamma)]), u \in D$. This provides a canonical description which is analogous to the one provided by Woods [6] in the context of Gauss Markov random fields on lattices.

References

[1] Yu. A. Rozanov, *Markov Random Fields.* New York: Springer-Verlag, 1982.

[2] T. Hida, *Brownian Motion.* New York: Springer-Verlag, 1980.

[3] L.D.Pitt, "A Markov property for Gaussian processes with a multidimensional parameter," *Arch. Ration. Mech. Anal,* vol. 43, pt.4, pp.367-391, 1971

[4] J. Diestel and J. J. Uhl, *Vector Measures.* Providence, Rhode Island: American Mathematical Society, 1977.

[5] G. Kallianpur and V. Mandrekar, " The Markov property for generalized random fields," *Ann. Inst. Fourier. Grenoble,* vol. 24, pt.2, pp. 143-167, 1974.

[6] J. W. Woods, "Two-dimensional discrete Markovian fields," *IEEE Trans. Inform. Theory,* vol.18, pp.232-240, 1972.

[1] This work was partially supported by an ONR Grant # N00014-91-J1001

Non-Parametric Discriminatory Power

Hilary J. Holz, Murray H. Loew
Department of Electrical Engineering & Computer Science
The George Washington University, Washington, D.C. 20052

ABSTRACT

Discriminatory power is the relative usefulness of a feature for classification. Traditionally, feature-selection techniques have defined discriminatory power in terms of a particular classifier. Non-parametric discriminatory power allows feature selection to be based on the structure of the data rather than on the requirements of any one classifier. In previous research, we have defined a metric for non-parametric discriminatory power called relative feature importance (RFI) *[1]. In this work, we explore the construction of RFI through closed-form analysis and experimentation. The behavior of RFI is also compared to traditional techniques.*

Relative feature importance ranks features based on an estimate of their *relative potential* for class separation. A set of optimal, orthogonal features is *extracted* from each possible feature subset, in order to estimate the potential for separation contained in the subset. The separation between class-conditional joint feature distributions is measured in the transformed space. The contribution of each original feature to the separation in the transformed space is estimated. Note that the separation contributed is *relative* in the sense that the use of the other features in the subset is taken into account.

Because the features may not be independent, the method first must determine the *optimal* subset of features. The optimal subset of original features is the smallest subset that yields maximal separation in the transformed space. Features outside the optimal subset are assigned an RFI of zero. The features within the optimal subset are ordered by their estimated contribution. The rank of a feature is its RFI.

Some critical design choices for RFI are: the feature extraction technique, the measure of separation in the transformed space, and the technique used to estimate the contribution of the original features to separation in the transformed space.

Rather than calculating RFI based on separation between the class-conditional joint feature distributions of the original features, the method uses a non-parametric feature extraction technique and calculates separation in the transformed space. Our extraction technique is based on Fukunaga and Mantock's non-parametric discriminant analysis [2]. Briefly, the data is expanded in the eigenvectors of the ratio of the within-class to the between-class scatter matrices. RFI uses non-parametric variations of within-class and between-class scatter matrices (which use local k-nearest-neighbor density estimates) and differentially weights samples based on their distance from the discriminant boundary. Since many traditional feature-ranking techniques are based on the marginal distributions of the original features, our examples include several multi-cluster experiments which cannot be solved using the marginals, but can be solved using extraction.

The contribution of the original features to separation in the transformed space is estimated using the Weighted Absolute Weight Size (WAWS) [1]. WAWS combines information from the magnitudes of the eigenvectors (which measure the contribution of the original features to the extracted features) and the normalized eigenvalues (which measure the amount of separation in the transformed space contributed by each extracted feature).

Through closed-form analysis and a series of experiments, we explore several design choices in the non-parametric feature extraction algorithm. We compare the algorithm's performance for Euclidean vs. Mahalanobis distance calculations, and for parametric vs. non-parametric scatter matrices for both within-class and between-class scatter. We consider several algorithms for calculating the non-parametric scatter matrices and for measuring separation in the transformed space. Each variant of the algorithm is evaluated according to several criteria.

A metric for non-parametric discriminatory power is important for a number of reasons. First, applications exist where optimizing the performance of an artificial classifier is the final goal: physicians need to know which test is the most accurate predictor of a particular disease whether or not they wish to use a classifier system as a diagnostic aid. Second, non-parametric discriminatory power can be used to direct feature search, without first having to select a classifier. Finally, a classifier, when desired, can be selected based on the distributional structure of the high-ranking features. All of these activities are currently undertaken in an ad-hoc fashion by humans using mapping and projection techniques. Unfortunately, such techniques are of limited utility in high-dimensional spaces.

[1] H.J. Holz and M.H. Loew, "Relative Feature Importance: Towards A Classifier-Independent Approach to Feature Selection," *Pattern Recognition in Practice IV,* Elsevier Sci. Pub., in press.

[2] K. Fukunaga and J. Mantock, "Nonparametric discriminant analysis," *IEEE Trans. Pattern Anal. Machine Intell.,* PAMI-5, pp. 671-178, Nov. 1983.

Shannon-Hartley Entropy Ratio under Zipf Law

R.E.Krichevskii, *Member, IEEE*, and M.P.Scharova
Math. Institute and State University, 630090 Novosibirsk, Russia

Abstract — **We find a formula for Shannon's-Hartley's Entropy Ratio of a text governed by the Zipf law. The formula is in a good agreement with real texts. It is asymptotically**

$$1/2 + \frac{loglog|D|}{log|D|},$$

$|D|$ being the size of the text's dictionary, $\lim|D| = \infty$.It means that a word to variable length code might significantly outperform a word to fixed length one for large dictionaries only.

I. Theoretical Considerations

Let T be a text, D be the set of all words of T (dictionary),$p(i)$ be the frequency of the i-th word, $i = 1, ..., |D|$.According to the Zipf law,

$$p(i) = A/i, \tag{1}$$

$$A = const, i = 1, ..., |D|.$$

Obviously,

$$A\sum_{i=1}^{|D|}\frac{1}{i} = 1 \tag{2}$$

. The well-known formula for the harmonic sum and (2) yield

$$A = \frac{1}{ln|D| + c_1} \tag{3}$$

where $c_1 = 0,577...$ is the Euler constant. Per word Hartley entropy of T equals $\log|D|$. It is the cost of a per word fixed length encoding of T (to within an additive constant). Per word Shannon entropy of T equals

$$H = -\sum_{i=1}^{n} p(i)logp(i) \tag{4}$$

. It is the cost of a per word variable length encoding of$T$ (to within an additive constant). By the Euler-Maclaurin relation between sums and integrals we obtain

$$\sum_{1=1}^{|D|}\frac{lni}{i} = \frac{ln^2|D|}{2} + c_2 + o(1), \tag{5}$$

where $c_2 = 0,211...$ ¿From (1) and (3)-(5) we get for the ratio of entropies

$$\frac{H}{loq|D|} = (\frac{1}{2}(1+\frac{c_1}{ln|D|})^{-1} + \frac{ln(ln|D| + c_1)}{ln|D|} + \frac{c_2 ln2}{ln^2|D| + c_1 ln|D|}) \tag{6}$$

If $\lim\log|D| = \infty$, then

$$\frac{H}{\log|D|} = \frac{1}{2} + \frac{lnln|D|}{\ln|D|} \tag{7}$$

II. Experiments

We have the following experimental results up to now.

1.Collected works of Russian poet A.S.Pouschkin. $|D| =$ 21197 words, Theoretical Ratio $\frac{H}{log|D|} =$ 0,710. Real Ratio $\frac{H}{log|D|} =$ 0,726.

2.The book of R.E.Krichevskii "Universal Compression and Retrieval", Kluwer Publishers. $|D| = 1900$ words, Theoretical Ratio $\frac{H}{log|D|} = 0,78$. Real Ratio $\frac{H}{log|D|} = 0,80$.

3.The book of Y.G.Reschetnjak "Space Mappings with Bounded Distortion", Providence, R.I. $|D| = 2000$ words Theoretical Ratio $\frac{H}{log|D|} = 0,78$. Real Ratio $\frac{H}{log|D|} = 0,77$.

4.A paper from Siberian Mathematical Journal. $|D| = 2100$ words, Theoretical Ratio $\frac{H}{log|D|} = 0,78$. Real Ratio $\frac{H}{log|D|} = 0,79$.

5.Another paper from Siberian Mathematical Journal. $|D| = 1600$ words, Theoretical Ratio $\frac{H}{log|D|} = 0,79$. Real Ratio $\frac{H}{log|D|} = 0,76$.

Mismatched Encoding in Rate Distortion Theory

Amos Lapidoth[1]

Information Systems Laboratory, Stanford Universty, Stanford, CA, 94305-4055.

Abstract —

A length n block code $\mathcal{C}$ of size 2^{nR} over a finite alphabet $\hat{\mathcal{X}}_0$ is used to encode a memoryless source over a finite alphabet $\mathcal{X}$. A length n source sequence x is described by the index i of the codeword $\hat{\mathbf{x}}_0(i)$ that is nearest to x according to the single-letter distortion function $d_0(x, \hat{x}_0)$. Based on the description i and the knowledge of the codebook $\mathcal{C}$, we wish to reconstruct the source sequence so as to minimize the average distortion defined by the distortion function $d_1(x, \hat{x}_1)$, where $d_1(x, \hat{x}_1)$ is in general different from $d_0(x, \hat{x}_0)$. In fact, the reconstruction alphabets $\hat{\mathcal{X}}_0$ and $\hat{\mathcal{X}}_1$ could be different.

We study the minimum, over all codebooks $\mathcal{C}$, of the average distortion between the reconstructed sequence $\hat{\mathbf{x}}_1(i)$ and the source sequence x as the blocklength n tends to infinity. This limit is a function of the code rate R, the source's probability law, and the two distortion measures $d_0(x, \hat{x}_0)$, and $d_1(x, \hat{x}_1)$.

This problem is the rate-distortion dual of the problem of determining the capacity of a memoryless channel under a possibly suboptimal decoding rule.

The performance of a random i.i.d. codebook is found, and it is shown that the performance of the "average" codebook is in general suboptimal. The resulting distortion can in general be improved by considering i.i.d. codebooks of m-tuples. It is shown that as m tends to infinity, the performance of the "average" codebook becomes optimal.

By studying the special case of a Gaussian source and minimum Euclidean Distance description, i.e. $d_0(x, \hat{x}_0) = (x - \hat{x}_0)^2$, we obtain an improved upper bound on the rate distortion function for a Gaussian source and an arbitrary distortion measure.

By exploring the analogy between the rate distortion problem and the mismatched channel decoding problem, we find that for an i.i.d. real-valued source of second moment σ^2, a random Gaussian codebook of size 2^{nR} achieves, for sufficiently large n, an average mean-square-error distortion of $2^{-2R}\sigma^2$, irrespective of the source distribution.

ACKNOWLEDGEMENTS

Stimulating discussions with Tom Cover, Aaron Wyner, and Emre Telatar are gratefully acknowledged.

[1]This research has been carried out in part while the author was with AT&T Bell Laboratories, Murray Hill, NJ.

SNR Estimation and Blind Equalization (Deconvolution) Using the Kurtosis

Rolf Matzner and Klemens Letsch

Federal Armed Forces University Munich, Inst. for Commun. Eng. ET 3, 85577 Neubiberg, Germany

Abstract — **The underlying mathematical problem of both, SNR estimation and blind equalization, is the sum of random processes. It can be shown that it is sufficient to describe the random processes as well as their sum by a shape factor of the p.d.f., the kurtosis, which includes the second and fourth order moment.**

I. Introduction

In the following we concentrate on discrete time signals and systems, although the algorithms in principle are applicable to analog systems as well.

A discrete time random process is a sequence of identically distributed random variables (r.v.) $\mathbf{x}_\mu$. If the random process is complex, each r.v. consists of real and imaginary part $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i$ with joint p.d.f. $f_x(x_r, x_i)$.

II. Formulation of the Problem

The *SNR estimation* problem can be described as follows: Given a wanted signal random process $\langle \mathbf{r} \rangle$ with (joint) p.d.f. $f_r(r_r, r_i)$ and a (statistically independent) noise random process $\langle \mathbf{n} \rangle$ with p.d.f. $f_n(n_r, n_i)$, estimate the signal-to-noise ratio $SNR = S_r/S_n$ just by observation of the sum process $\langle \mathbf{y} \rangle = \langle \mathbf{r} \rangle + \langle \mathbf{n} \rangle$. Usually the type of p.d.f. of the wanted signal (e.g. M-PAM, M-PSK, etc.) and the noise (e.g. Gaussian) are known, while in general neither received signal power nor noise power are known.

Blind equalization means to find a filter

$$\langle c \rangle = c_0, c_1, \ldots, c_{N_c-1}$$

for an unknown system (channel) $\langle h \rangle$ such that the overall impulse response $\langle s \rangle = \langle c \rangle \circ \langle h \rangle = 1$, where "$\circ$" denotes discrete convolution. "Blind" means that only the output $\langle z \rangle = \langle x \rangle \circ \langle s \rangle$ is observable, while the input sequence $\langle x \rangle$ is unknown. Of course the knowledge of some statistical properties of $\langle \mathbf{x} \rangle$ is necessary. More specifically we demand that $\langle \mathbf{x} \rangle$ be an i.i.d. sequence with known kurtosis.

Hence the output random process is a weighted sum of a number of i.i.d. input random processes $\langle \mathbf{x} \rangle_k$:

$$\langle \mathbf{z} \rangle = \sum_k s_k \langle \mathbf{x} \rangle_k \tag{1}$$

The parallelism to the SNR estimation problem is obvious.

III. The Sum of Random Processes (Variables) and the Kurtosis

We assume all input random processes to be complex valued, stationary and zero-mean with existing characteristic function and moments up to the fourth order. Moreover, the channel input $\langle \mathbf{x} \rangle$ must be i.i.d..

The $(k+\ell)$-th order moment of a complex random process $\langle \mathbf{r} \rangle$ is defined by

$$m_r(k,\ell) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} r_r^k r_i^\ell f_r(r_r, r_i) dr_r dr_i. \tag{2}$$

It can be shown that the $(k+\ell)$-th order moment of the sum $\langle \mathbf{y} \rangle = \langle \mathbf{r} \rangle + \langle \mathbf{n} \rangle$ can be expressed as

$$m_y(k,\ell) = \sum_{u=0}^{k}\sum_{v=0}^{\ell} \binom{k}{u}\binom{\ell}{v} m_r(k-u, \ell-v) m_n(u,v). \tag{3}$$

We define the kurtosis of a random process $\langle \mathbf{r} \rangle$ as he ratio of the fourth order to the squared second order expected value

$$K_r = E\{rrr^*r^*\} / \left(E\{rr^*\}\right)^2. \tag{4}$$

Replacing expected values by moments gives:

$$K_r = \frac{m_r(4,0) + 2m_r(2,2) + m_r(0,4)}{\left(m_r(2,0) + m_r(0,2)\right)^2}. \tag{5}$$

Using (2), (3) and (5) we can express the kurtosis K_y of $\langle \mathbf{y} \rangle$ as:

$$K_y - K_G = \kappa^2 (K_r - K_G) + (1-\kappa)^2 (K_n - K_G), \tag{6}$$

where $\kappa = S/(S+N)$ is the wanted signal power to total received power ratio and K_G is the kurtosis of a Gaussian process. Eq. (6) is the motivation to define the *Gauss unlikeness* $G_r = K_r - K_G$, cf. [1]

IV. SNR Estimation

The algorithm to estimate the SNR is quite simple now. G_r and G_n are known from the used modulation scheme and the expected noise type, resp. Then observe the received signal $\langle \mathbf{y} \rangle$ and compute an estimate for G_y by averaging in the time domain. Finally solve (6) for κ and the estimated SNR is $SNR = \kappa/(1-\kappa)$.

V. Blind Equalization

Extending (6), the Gauss unlikeness G_z at the output of the equalizer is given by

$$G_z = G_x \sum_k s_k^4 \Big/ \left(\sum_k s_k^2\right)^2. \tag{7}$$

$|G_z|$ is always less or equal to $|G_x|$, with equality if and only if either $\langle \mathbf{x} \rangle$ is Gaussian or $\langle s \rangle = 1$. The former case cannot be solved by any means, and the latter describes a perfectly equalized channel. Hence it is sufficient to maximize $|G_z|$, e.g. using a stochastic gradient algorithm (see [2]).

References

[1] R. Matzner and F. Englberger, "An SNR estimation algorithm using fourth-order moments," in *Int. Symp. Inf. Theory*, (Trondheim), 1994.

[2] R. Matzner and A. Schmidbauer, "Blind linear and decision feedback equalizers using fourth-order moments and their performance on twisted pair lines," in *Proc. GLOBECOM '94*, (San Francisco, CA), Nov. 1994. (to appear).

Minimum Complexity Regression Estimation With Weakly Dependent Observations

Dharmendra S. Modha and Elias Masry[1]

Department of Electrical & Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407

Abstract — **Given N strongly mixing observations $\{X_i, Y_i\}_{i=1}^N$, we estimate the regression function $f^*(x) = E[Y_1|X_1 = x]$, $x \in \Re^d$ from a class of neural networks, using certain minimum complexity regression estimation schemes. We establish a rate of convergence for the integrated mean squared error between the proposed regression estimator and f^*.**

I. Introduction

Let $\{X_i, Y_i\}_{i=-\infty}^{\infty}$ be a stationary process such that X_1 takes values in $\Re^d$ and Y_1 takes values in $\Re$. Given N observations $\{X_i, Y_i\}_{i=1}^N$ drawn from $\{X_i, Y_i\}_{i=-\infty}^{\infty}$, we are interested in postulating an estimator based on single hidden layer sigmoidal networks for the regression function $f^* = E[Y_1|X_1 = x]$, $x \in \Re^d$.

Recently, assuming that the underlying random variables $\{X_i, Y_i\}_{i=-\infty}^{\infty}$ are i.i.d., Barron [1] proposed a minimum complexity regression estimator based on single hidden layer sigmoidal networks. Moreover, supposing that Assumption 1 (see below) holds he established a rate of convergence for the integrated mean squared error between his estimator and f^*. In this paper, we extend Barron's results from i.i.d. random variables to stationary strongly mixing [3] processes. The reader is referred to the full paper [2] for complete analysis.

II. A Class of Target Regression Functions and Single Hidden Layer Sigmoidal Networks

Assumption **1.** *Assume that* (a) Y_1 *takes values in some interval* $\mathcal{I} \equiv [a, a+b] \subset \Re$ *a.s.;* (b) X_1 *takes values in* $\mathcal{B} \equiv [-1,1]^d$ *a.s.; and that* (c) *there exists a complex valued function* $\tilde{f}$ *on* $\Re^d$ *such that for* $x \in \mathcal{B}$, *we have*

$$f^*(x) - f^*(0) = \int_{\Re^d} \left(e^{iw\cdot x} - 1\right) \tilde{f}(w)\, dw$$

and that $\int_{\Re^d} ||w||_1 |\tilde{f}(w)|\, dw \leq C' < \infty$ *for some known* $C' > 0$. *Set* $C = \max\{1, C'\}$.

Let $\phi : \Re \to \Re$ denote a sigmoidal function such that $|\phi(u) - 1_{\{u>0\}}| \leq q'/|u|^p$ for some $p > 0$, $q' \geq 0$, and for all $u \in \Re \setminus \{0\}$. Set $q = \max\{1, q'\}$. For $n \geq 1$, let $\gamma_n = n(d+2)+1$. For $0 \leq i \leq n$, let $c_i \in \Re$; for $1 \leq i \leq n$, let $a_i \in \Re^d$ and let $b_i \in \Re$. We define a γ_n-dimensional parameter vector $\theta^{(n)}$ as

$$\theta^{(n)} = (a_1, a_2, \ldots, a_n;\ b_1, b_2, \ldots, b_n;\ c_0, c_1, \ldots, c_n).$$

Now, define a single hidden layer sigmoidal network $f_{\theta^{(n)}} : \Re^d \to \Re$ parametrized by $\theta^{(n)}$ as

$$f_{\theta^{(n)}}(x) = c_0 + \sum_{i=1}^{n} c_i\, \phi(a_i \cdot x + b_i), \quad x \in \Re^d. \tag{1}$$

[1]This work was supported by the Office of Naval Research under Grant N00014-90-J-1175.

Set $\varpi_n = 2^{\frac{2p+1}{p}}\, q^{\frac{1}{p}}\, n^{\frac{p+1}{2p}}$ and define $\mathcal{S}^{(n)} \subset \Re^{\gamma_n}$ as

$$\{\theta^{(n)} : c_0 \in \mathcal{I}, \sum_{i=1}^{n} |c_i| \leq 2C, \max_{1\leq i\leq n} ||a_i||_1 \leq \varpi_n, \max_{1\leq i\leq n} |b_i| \leq \varpi_n\}.$$

For each fixed n and N and given an $\epsilon_{n,N} > 0$, we construct an $\epsilon_{n,N}$-net of $\mathcal{S}^{(n)}$, namely, $T_{n,N}$ such that

$$\ln \text{card}(T_{n,N}) \leq \gamma_n \ln \frac{4\varpi_n e}{\epsilon_{n,N}} \equiv L_{n,N},$$

where $\text{card}(T_{n,N})$ denotes the cardinality of the set $T_{n,N}$.

III. Estimation Scheme and Main Result

Let $\alpha(j)$ denote the strong mixing coefficient [3] corresponding to the process $\{X_i, Y_i\}_{i=-\infty}^{\infty}$.

Assumption **2.** *Assume that the strong mixing coefficient satisfies* $\alpha(j) = \bar{\alpha}\exp(-cj^{\beta}), j \geq 1, \bar{\alpha} \in (0,1], \beta > 0, c > 0$.

Write $l_N = \lfloor N\lceil\{8N/c\}^{1/(\beta+1)}\rceil^{-1}\rfloor$. l_N plays the same role in our analysis as the sample size N in the i.i.d. case. Define

$$\hat{\theta}_{n,N} = \underset{\theta\in T_{n,N}}{\arg\min} \left\{\frac{1}{N}\sum_{i=1}^{N}(Y_i - f_\theta(X_i))^2\right\},$$

where for a given $\theta \in T_{n,N}$, f_θ is defined as in (1). Now, for each fixed regularization constant $\lambda > 0$, define $\hat{n} \equiv \hat{n}_N$ as

$$\underset{1\leq n\leq l_N}{\arg\min} \left\{\frac{1}{N}\sum_{i=1}^{N}(Y_i - f_{\hat{\theta}_{n,N}}(X_i))^2 + \lambda\frac{L_{n,N} + 2\ln(n+1)}{l_N}\right\},$$

and define the *minimum complexity estimator* as $f_{\hat{\theta}_{\hat{n},N}}$.

Theorem **1.** *Suppose Assumptions 1 and 2 hold. Let* $\lambda > 5b^2/3$ *and for some* $r \geq 1/2$ *let* $(nl_N)^{-r} \leq \epsilon_{n,N} \leq n^{-1/2}$, *then*

$$E\int_{\Re^d}[f_{\hat{\theta}_{\hat{n},N}}(x) - f^*(x)]^2\, dP_X(x) = O\left(\frac{\sqrt{\ln N}}{N^{\beta/(2\beta+2)}}\right), \tag{2}$$

where P_X *denotes the marginal distribution of* X_1.

Note that the exponent of N in (2) does not depend on the dimension d. In [2], we compare the rate of convergence obtained in Theorem 1 to the rate of convergence achieved by the classical nonparametric kernel estimator in similar setting and to the rate of convergence obtained by Barron [1] in the i.i.d setting. In [2], we also establish a result analogous to Theorem 1 for m-dependent observations.

References

[1] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, pp. 115-133, 1994.

[2] D. S. Modha and E. Masry, "Minimum complexity regression estimation with weakly dependent observations," submitted for publication, 1994.

[3] M. Rosenblatt, "A central limit theorem and strong mixing conditions," *Proc. Nat. Acad. Sci.*, vol. 4, pp. 43-47, 1956.

EM and SAGE Algorithms for Multi-user Detection

Laurie B. Nelson H. Vincent Poor
laurie@ee.princeton.edu poor@princeton.edu
Dept. of Electrical Engineering, Princeton University, Princeton, New Jersey 08544

Abstract — **This work describes an EM-based approach to multi-user detection that treats the signals of interfering users as hidden data. We consider a new algorithm based on the *space-alternating generalized EM (SAGE) algorithm* appropriate for estimation of discrete random parameters, and we use it to derive rapidly-convergent nearly optimum multi-user receivers.**

I. Introduction

The *expectation-maximization (EM) algorithm* [1] provides an iterative approach to parameter estimation when direct maximization of the likelihood function may be infeasible. The recently proposed SAGE algorithm [2] modifies EM to update only a subset of the parameter components at each iteration, thereby allowing the use of less-informative hidden data in order to improve convergence rates. We consider a new algorithm based on the SAGE structure that incorporates the statistics of the parameter components not currently being updated. Our motivation is the problem of multi-user detection [3], for which the vector parameter $\mathbf{b}$ corresponds to the binary data of several users in a CDMA system. The complexity of optimum decisions for $\mathbf{b}$ (under either Bayesian or ML criteria) is exponential in the number of users [3] and motivates the development of simpler iterative receivers.

Modifying the SAGE algorithm to treat parameter components not currently being updated as hidden data leads to a new "hidden-parameter" EM (HPEM) algorithm. In terms of the observation $\mathbf{Y}$ and vector parameter $\mathbf{b}$ with joint density $f(\mathbf{y},\mathbf{b})$, the ith iteration of the HPEM algorithm is described by the following steps:

- Set $k = 1 + (i \bmod K)$. Let $\mathbf{b}_{\tilde{k}} = \{b_j : j \neq k\}$. Choose the hidden data $\mathbf{x}$.
- E-step: Compute

$$Q(b_k;\mathbf{b}^i) \triangleq \int \log f(\mathbf{y},\mathbf{x},\mathbf{b}_{\tilde{k}} \mid b_k)\, h(\mathbf{x},\mathbf{b}_{\tilde{k}};\mathbf{y},\mathbf{b}^i)\, d\mathbf{x} d\mathbf{b}_{\tilde{k}} \quad (1)$$

- M-step: $b_k^{i+1} = \arg\max_{b_k} Q(b_k;\mathbf{b}^i)$, $\mathbf{b}_{\tilde{k}}^{i+1} = \mathbf{b}_{\tilde{k}}^i$.

In (1), the integrating density h is given by either the conditional density $f(\mathbf{x},\mathbf{b}_{\tilde{k}} \mid \mathbf{y}, b_k)$ or the product of conditional densities

$$f(\mathbf{x} \mid \mathbf{y},\mathbf{b}^i) \textstyle\prod_{m\neq k} f(b_m \mid \mathbf{y}, \mathbf{b}_{\tilde{m}}^i). \quad (2)$$

For the former case, we have

$$Q(b_k;\mathbf{b}^i) = E\left\{\log f(\mathbf{y},\mathbf{x},\mathbf{b}_{\tilde{k}} \mid b_k) \mid \mathbf{y}, b_k = b_k^i\right\},$$

which is essentially a smoothed version of the SAGE E-step objective function. This algorithm produces an estimate sequence that is non-decreasing in the *marginal* likelihoods

$$\log f(\mathbf{y} \mid b_k = b_k^{i+1}) \geq \log f(\mathbf{y} \mid b_k = b_k^i) \quad (3)$$

for all $k = 1,\ldots,K$, where $f(\mathbf{y} \mid b_k) = E\{f(\mathbf{y},\mathbf{b} \mid b_k))\}$.

This research was supported by the U.S. Army Research Office under Grant DAAH04-93-G-0219 and by an AT&T Ph.D. Fellowship.

Intuitively, the latter case (2) involves conditioning the statistics of each hidden parameter b_m on current estimates for *all* parameter components except b_m, rather than just on $b_k = b_k^i$. Under mild conditional independence assumptions, the algorithm with h given by (2) also produces an estimate sequence satisfying (3).

II. Receivers for "Hidden Interferers"

The received signal in the K-user synchronous CDMA channel is described by $\mathbf{Y} \sim N(\mathbf{RAb}, \sigma^2\mathbf{R})$, where $\mathbf{R}$ is a positive-definite, symmetric matrix of signature waveform cross-correlations, $\mathbf{A}$ is a diagonal matrix of the users' signal amplitudes, and $\mathbf{b} \in \{\pm 1\}^K$ is the users' data over one bit interval. The observation $\mathbf{Y}$ corresponds to the sampled outputs of filters matched to the users' signature waveforms.

In applying the HPEM algorithm to multi-user detection, we assume $\mathbf{b}$ is distributed equiprobably on $\{\pm 1\}^K$ and consider the scenario when $\mathbf{R}$, $\mathbf{A}$, and σ^2 are known. The resulting receiver cyclically updates estimates for the K users' bits. With respect to iterations that update b_k, the E-step computes soft-decision estimates of the interference:

$$\tilde{b}_m = E\left\{b_m \mid \mathbf{y}, \mathbf{b}_{\tilde{m}} = \mathbf{b}_{\tilde{m}}^i\right\} = \tanh\left(\frac{a_m}{\sigma^2}\left(y_m - \sum_{j\neq m} R_{mj} a_j b_j^i\right)\right)$$

for all $m \neq k$. The M-step cancels the estimated interference; i.e.,

$$z_k = y_k - \textstyle\sum_{m\neq k} R_{km} a_m \tilde{b}_m,$$

and updates the estimate for b_k via $b_k^{i+1} = \operatorname{sgn}(z_k)$. Alternatively, one could model the parameter b_k as taking values in the interval $[-c, c]$. In this case, the M-step update is

$$b_k^{i+1} = \begin{cases} c\operatorname{sgn}(z_k/a_k) & z_k/a_k \notin [-c,c] \\ z_k/a_k & z_k/a_k \in [-c,c]. \end{cases}$$

The HPEM receiver has a structure similar to multi-stage receivers [4], but it enjoys some unique and significant convergence and performance advantages due to the non-decreasing likelihood of the estimate sequence $\mathbf{b}^i$. These are verified by theory and simulation.

One might also consider application of the SAGE algorithm to multi-user detection. Depending on the M-step update nonlinearity, the resulting receivers provide either an iterative implementation of the decorrelating detector [3] or a convergent multi-stage receiver using sequential, rather than simultaneous, updates of the bit estimates.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J Roy Stat Soc, Ser B*, vol. 39, no. 1, pp. 1–38, 1977.

[2] J. A. Fessler and A. O. Hero, "Space-alternating generalized EM algorithm," *IEEE Trans. Signal Proc.*, Oct. 1994.

[3] S. Verdú, "Multiuser detection," *Advances in Statistical Signal Processing*, vol. 2, pp. 369–409, 1993.

[4] M. K. Varanasi and B. Aazhang, "Near-optimum detection in synchronous CDMA systems," *IEEE Trans. Comm.*, vol. 39, no. 5, pp. 725–736, May 1991.

Nonparametric Estimation of a Class of Smooth Functions

M.Pawlak and U.Stadtmüller
Dept. of Electrical and Computer Eng.,Univ.of Manitoba, Winnipeg, Canada R3T5V6,
e-mail:Pawlak@ee.umanitoba.ca
Abt. für Mathematik III, Univ. of Ulm, Helmholtzstr.18, D89069 Ulm, Germany

Abstract — The problem of recovering band-limited signals from noisy data is considered. Whittaker-Shannon cardinal expansions based estimates involving sampling windows and truncation of higher frequencies are introduced. Weak and strong pointwise convergence properties of the proposed estimates are derived.

I. Introduction

Consider the problem of recovering a function f(t), when only the noisy measurements generated by the following model are available $y_k = f(k\tau) + \varepsilon_k$, $k = 0, \pm1, \pm2, ...$, some $\tau > 0$, and ε_k represents noise.
The objective of this paper is to examine the statistical properties of reconstruction schemes for f(t) which is band-limited, i.e., which Fourier transform has support within $(-\Omega, \Omega)$, Ω is a finite number called the bandwidth of f. Such a function can be represented by the so-called cardinal series due originally to Whittaker and Shannon

$$f(t) = \sum_{k=-\infty}^{\infty} f(k\tau)\, \mathrm{sinc}(\frac{\pi}{\tau}(t - k\tau)), \qquad (1)$$

uniformly in any bounded interval of R, provided that $\tau \le \pi/\Omega$, where $\mathrm{sinc}(x) = \sin x/x$.

II. Estimation Techniques

The representation in (1) forms a natural basis for our estimation techniques. We construct a class of recovering algorithms of the following form

$$\hat{f}_\psi(t;\tau,\delta) = \tau \sum_{|j|\le n} y_j K_\delta(t - j\tau), \qquad (2)$$

where $K_\delta(t) = \frac{\sin((\delta+1)\Omega t)}{\pi t}\psi(\delta\Omega t)$, $0<\tau<\pi/\Omega$, $\delta>0$, ψ is a band-limited function with the bandwidth equals 1 and $\psi(0)=1$. For $\delta \to 0$ the estimate in (2) takes the form of the kernel convolution estimate with the kernel $\sin(\Omega t)/\pi t$. This clearly represents a truncated and smoothed version of the expansion in (1).
The pointwise properties of the proposed estimates are established which includes the consistency and convergence rate. In particular, it is shown that for a certain choice of the parameters τ, δ and the function ψ the mean squared difference between $\hat{f}_\psi(t;\tau,\delta)$ and f(t) can tend to zero as fast as $O(\frac{\ln n}{n})$ uniformly in t and for any f in the band-limited class.

References

[1] S. Cambanis and E. Masry, *Zakai's class of bandlimited functions and processes: its characterization and properties*, SIAM J. Appl. Math., **30** (1976), 10- 21.
[2] W. Härdle, *Applied Nonparametric Regression* , Cambridge University Press, Cambridge, 1990.
[3] J.J. Knab, *The sampling window*, IEEE Trans. Information Theory **29** (1983), pp.157-159.
[4] R.J. Marks II (ed.), *Advances Topics in Shannon Sampling and Interpolation Theory*, Springer -Verlag, New York, 1993.
[5] M. Pawlak and U. Stadtmüller, *Recovering band-limited functions under noise*, Report, 1994.
[6] A.I. Zayed, *Advances in Shannon's Sampling Theory*, CRC Press, Boca Raton, 1993.

Consistency and Rates of Convergence of k_n Nearest Neighbor Estimation Under Arbitrary Sampling[1]

S.E. Posner S.R. Kulkarni

Department of Electrical Engineering, Princeton University, Princeton, NJ 08544

Abstract — **Consistency and rates of convergence of the k_n-NN estimator are established in the general case in which samples are chosen *arbitrarily* from a compact metric space.**

I. Introduction

Nearest neighbor (NN) estimation has received much attention since it was studied in [1]. Most existing work generally considers the case in which the observations are drawn i.i.d. from a probability distribution. Some authors (e.g., [2, 4]) have discussed rates of convergence of k_n-NN estimators. In [3], we formulated a new estimation problem in which the observations need not be drawn randomly but can be *arbitrarily* selected. We investigated convergence of the NN rule in this framework and found its convergence rate. In this paper, we consider the consistency of the k_n-NN estimator under arbitrary sampling.

II. Problem Formulation and Preliminaries

Let $Y = \mathbb{R}^s$ (for some positive integer s) with inner-product induced norm $\|\cdot\|$ and let X be a metric space with metric ρ which we denote (X, ρ). Given a point $x_n \in X$, a random variable y_n is drawn with unknown conditional probability distribution $F(y_n|x_n)$. We are asked to produce an estimate, $\hat{y}_n$, of the value of y_n with the goal of minimizing $\|y_n - \hat{y}_n\|^2$. If $F(y_n|x_n)$ is known, the estimate that gives the minimum possible expected loss is the Bayes estimate, $r^*(x_n)$. If x_n is chosen arbitrarily, the minimum worst possible expected loss is the sup over x_n of the conditional Bayes risk, R^{w*}.

We consider the problem in which the only knowledge of $F(y|x)$ is that inferred from pairs of samples $(x_1, y_1), \cdots, (x_{n-1}, y_{n-1})$. The k_n-nearest neighbor rule is defined as follows. Let k_n be any nondecreasing sequence of numbers. Denote the k_n nearest neighbors of x_n as $x_n^{[1]}, \ldots, x_n^{[k_n]}$ where $x_n^{[1]}$ is the nearest. We denote $y_n^{[i]}$ as the parameter associated with the ith nearest neighbor. The k_n-NN rule estimate is the average of the k_n NN parameters, $\bar{y}_n^{(k_n)} = \frac{1}{k_n}\sum_{i=1}^{k_n} y_n^{[i]}$. Let $r_n^{(k_n)}(x_n, x_n^{[1]}, \ldots, x_n^{[k_n]})$ be the expected loss using the k_n nearest neighbors of x_n. If the x_i's are chosen arbitrarily, in general one cannot get a useful bound on $r_n^{(k_n)}(x_n, x_n^{[1]}, \ldots, x_n^{[k_n]})$. Instead we prove an upper bound on the cumulative risk. We define $C_n^{(k_n)}(x_1, \ldots, x_n)$ as the cumulative k_n-NN loss and $\bar{R}_n^{(k_n)}(x_1, \ldots, x_n)$ as the time-averaged risk of a given arbitrary sequence.

We impose the following Lipschitz assumption on $F(y|x)$. Let $m(x) = E[y|x]$ and $\sigma^2(x) = E[\|y\|^2|x] - \|m(x)\|^2$. Note that $r^*(x) = \sigma^2(x)$.

Assumption 1 *There exists $K, \alpha > 0$ such that for any $x_1, x_2 \in X$,*

$$\|m(x_1) - m(x_2)\| \le \sqrt{K}\rho(x_1, x_2)^\alpha$$

$$|\sigma^2(x_1) - \sigma^2(x_2)| \le K\rho(x_1, x_2)^{2\alpha}$$

The metric covering number $\mathcal{N}(\epsilon, A)$ of a compact subset A of (X, ρ) is defined as the smallest number of open balls of radius ϵ that cover A. The inverse function, $\mathcal{N}^{-1}(k, A)$, the metric covering radius, is the smallest radius such that there exists k balls of this radius that cover A.

III. Main Result

Theorem 1 *With squared error loss, for any $F(y|x)$ satisfying Assumption 1, and with any arbitrary sequence $x_1, \ldots, x_n$ in compact subset A of (X, ρ), we have that any k_n-nearest neighbor rule satisfies*

$$C_n^{(k_n)}(x_1, \ldots, x_n) \le \sum_{i=2}^{n}\left(1 + \frac{1}{k_i}\right) r^*(x_i) + 2K\sum_{i=k_n}^{n-1}[2\mathcal{N}^{-1}(\lfloor i/k_n \rfloor, A)]^{2\alpha}$$

If $\{k_n\}$ satisfies (C1) $k_n \to \infty$ and (C2) $\frac{k_n}{n} \to 0$ then (when the limit exists) we have that

$$\lim_{n\to\infty} \bar{R}_n^{(k_n)}(x_1, \ldots, x_n) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=2}^{n} r^*(x_i)$$

$$\lim_{n\to\infty} \sup_{x_1,\ldots,x_n} \bar{R}_n^{(k_n)}(x_1, \ldots, x_n) = R^{w*}$$

The first statement gives an upper bound on the cumulative risk for an arbitrary sequence in terms of the sum of the conditional Bayes risks plus the growth rate. The rate is independent of the sequence but is in terms of an intrinsic topological quantity of the compact set A. It quantifies how close arbitrary points in a compact set cluster together. The final statement states that the asymptotic time-average of the NN risk equals the time-average of the conditional risks of the particular sequence. Also, the asymptotic time-average of the worst possible sequence equals the worst possible conditional Bayes risk.

In particular, for compact subsets of $\mathbb{R}^r$, it is well-known that $\mathcal{N}^{-1}(n, A) = O(n^{-1/r})$. This gives a convergence rate of $O(n^{-\frac{2\alpha}{r+2\alpha}})$ for the time-averaged risk. This rate coincides with rates established [2] in the random sampling case.

[1] This work was supported in part by the National Science Foundation under grants IRI-9209577 and IRI-9457645 and by the U.S. Army Research Office under grant DAAL03-92-G-0320.

References

[1] T.M. Cover, "Estimation by the nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 50-55, 1968.

[2] L. Györfi, "The rate of convergence of k_n-NN regression estimates and classification rules," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 3, pp. 362-364, 1981.

[3] S.E. Posner and S.R. Kulkarni, "Rates of convergence of nearest neighbor estimation under arbitrary sampling," IEEE Information Theory Symposium, 1994.

[4] C.J. Stone, "Consistent nonparametric regression," *Ann. Stat.*, vol. 5, pp. 595-645, 1977.

Choosing Data Sets that Optimize the Determinant of the Fisher Information Matrix

Wendy L. Poston and Jeffrey L. Solka[1]
G33, NSWCDD, Dahlgren, VA 22448-5000

Abstract – **In many situations it is desirable to operate on a subset of the data only. These can arise in the areas of experimental design, robust estimation of multivariate location, and density estimation. This paper will describe a method of subset selection that optimizes the determinant of the Fisher Information Matrix (FIM) which is called the Effective Independence Distribution (EID) method. Some motivation will be provided that justifies the use of the EID, and the problem of finding the subset of points to use in the estimation of the Minimum Volume Ellipsoid (MVE) will be examined as an application of interest.**

I. Problem Statement

The determinant of the FIM as an objective function to optimize arises in many areas of statistics and engineering. In most cases, the problem is one of subset selection which can be stated as follows: given a data set of size n, select a subset of these points of size m, where $m<n$, such that the determinant of the FIM is optimized. It is assumed that each data point is dimension p and that $n>>p$. Subset selection should not be confused with dimensionality reduction, where the goal is to reduce the value p.

Current methods of subset selection typically rely on random methods [1]. The problem with these methods is that they are not guaranteed to find the global optimum with respect to the objective function in any finite sampling. Another undesirable aspect of these is that the results are not reproducible because they are based on randomly selected subsets.

II. Effective Independence Distribution

The EID was developed and used by Kammer [2] to choose optimal sensor locations for a modal identification experiment on the space station. The EID can be calculated as the diagonal elements of the following matrix

$$\text{EID} = \text{diag}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$$

where $\mathbf{X}$ is an $n \times p$ data matrix with each row containing one data point, and the FIM is given by

$$\text{FIM} = \mathbf{X}^T\mathbf{X}$$

It can be shown [3] that the following relationship holds between the determinants of the FIM as points are removed from a data set

$$\left|\mathbf{X}_{-i}^T\mathbf{X}_{-i}\right| = (1 - \text{EID}_i)\left|\mathbf{X}^T\mathbf{X}\right|$$

where $\mathbf{X}_{-i}$ is the data matrix with the *i-th* point removed and EID_i corresponds to the *i-th* data point. From this, it is apparent that the required optimization of the determinant of the FIM can be obtained by deleting observations with the appropriate EID value. Further theory motivating the use of the EID method will be provided in the poster session.

III. Application

The MVE estimator [1] is a robust estimator of location and covariate structure. Determining the MVE consists of two parts: 1) finding the subset of points to be used in the estimate and 2) finding the ellipse that covers this set. The EID method addresses the first problem.

To test the usefulness of the EID method, it is applied to six data sets where the true MVE is known. The paper by Hawkins [4] gives the correct subset and the resulting volume of the true MVE for these data sets. These are regression data, and the predictors are used to determine the MVE. The size of the data sets are relatively small, ranging from n=20 to n=50. The dimensionality of the data is $2 \le p \le 5$.

The EID algorithm is used to determine the set of points that comprise the MVE estimate. It is implemented in MATLAB on a 486, 33MHz computer, and the time required to find the subset of points ranges from 0.11 seconds to 0.77 seconds. The relative error in the volumes of the minimum covering ellipsoid using the EID approach are less than 6% for these data sets.

IV. Summary

In this paper, the EID method of subset selection has been described. Since this is a deterministic method, the results are repeatable for a given data set which is a desirable property. This method is relevant for a wide range of applications.

References

[1] P. J. Rousseeuw & A. M. Leroy, *Robust Regression and Outlier Detection,* John Wiley & Sons, 1987.

[2] D. C. Kammer, "Sensor placement for on-orbit modal identification and correlation of large space structures," *J. Guidance Control & Dynamics,* 1991.

[3] W. L. Poston & R. H. Tolson, "Maximizing the determinant of the information matrix with the effective independence distribution method," *J. Guidance Control & Dynamics,* 1992.

[4] D. M. Hawkins, "A feasible solution for the minimum volume ellipsoid estimator in multivariate data," *Comp Stats,* 1993.

[1] This work was supported by the NSWCDD Independent Research Program.

The Application of Akaike Information Criterion Based Pruning to Nonparametric Density Estimates

Jeff Solka[1], Carey Priebe[2], George Rogers[1], Wendy Poston[1],

and David Marchette[1]

[1]Naval Surface Warfare Center Dahlgren Division

Code B10

Dahlgren VA 22448-5000

[2] Department of Mathematical Sciences

The John Hopkins University

Baltimore Md. 21218

Abstract

This paper examines the application of Akaike Information Criterion (AIC) based pruning to the refinement of nonparametric density estimates obtained via the Adaptive Mixtures (AM) procedure of Priebe and Marchette. The paper details a new technique that uses these two methods in conjunction with one another to predict the appropriate number of terms in the mixture model of an unknown density. Results that detail the procedure's performance when applied to different distributional classes will be presented. Results will be presented on artificially generated data, well known data sets, and some recently collected features for mammographic screening.

I. APPROACH

Given $X=\{x_1, x_2, ..., x_n\}$ where each x_i is i.i.d. according to an unknown density $\alpha(x)$ then one is often interested in estimating $\alpha(x)$. This problem occurs in many areas. There are a variety of approaches to the multivariate density estimation problem [1].

An often used parametric approach is that of finite mixture models [2] in combination with the expectation maximization (EM) method of Dempster, Laird, and Rubin [3]. Given an unknown distribution $\alpha(x)$ we seek to model the distribution using $\alpha^*(x)$ defined by

$$\alpha^*(x;\Psi) = \sum_{i=1}^{m} \pi_i K(x;\Gamma_i) \qquad (1)$$

where K is some fixed density parameterized by Γ, and $\Psi = (\pi_1, \Gamma_1, \pi_2, \Gamma_2, ..., \pi_m, \Gamma_m)$. The π_i's are referred to as the mixing proportions. (We can assume for much of what follows that K is taken to be the univariate normal distribution, in which case Γ_i becomes $\{\mu_i, \sigma_i\}$.) One difficulty with the finite mixtures approach is that one needs some idea as to the appropriate number of terms in the mixture model.

A recently developed density estimation technique that circumvents some of the problems of the above technique is the adaptive mixtures density estimation (AMDE) procedure of Priebe [4]. This procedure is a blend of the finite mixtures and kernel estimator approach. It is essentially a mixtures type approaches that allows for the creation of new terms as indicated by the data complexity. It is important to note that unlike finite mixture models the number of terms m is not fixed but is estimated from the data. A problem with the adaptive mixtures procedure is that the solutions that it produces are typically overdetermined. While being good functional estimate of $\alpha(x)$ they have too many terms in the mixture.

Using the Akaike Information Criterion (AIC) [5] as a starting point we have developed a procedure that uses a single or set of adaptive mixtures density estimates and produces a set of pruned models with a lower complexity. This procedure attempts to obtain a minimum complexity model by iteratively pruning terms from the original model. The keys to this approach are AIC based pruning of AMDE models based on resampled data sets.

II. RESULTS

Results obtained using this procedure were presented at the poster session.

III. REFERENCES

[1] D. W. Scott, *Multivariate Density Estimation*, John Wiley and Sons, New York, NY, 1992.

[2] B. S. Everitt and D. J. Hand, *Finite Mixture Distributions*, Chapman and Hall. London, UK, 1981.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm" *J. Royal Statist. Soc., Series B* 39 1-38, 1977.

[4] C. E. Priebe, "Adaptive Mixtures", JASA vol 89, No. 427, pp 796-806, 1994.

[5] H. Akaike, "A New Look at the Statistical Model Identification" *IEEE Trans. Auto. Control*, vol.19, pp.716-723, 1974.

Improved Ziv-Zakai Lower Bound for Vector Parameter Estimation

Kristine L. Bell, Yossef Steinberg, Yariv Ephraim, and Harry L. Van Trees
ECE Dept., Center of Excellence in C3I, George Mason Univ., Fairfax, VA 22030-4444

I. Introduction

Ziv-Zakai bounds [1]-[4] on the mean-square-error (MSE) in parameter estimation are some of the tightest available bounds. These bounds relate the MSE in the estimation problem to the probability of error in a binary hypothesis testing problem. The original Bayesian version derived by Ziv and Zakai [1], and improvements by Chazan-Zakai-Ziv [2] and Bellini-Tartara [3] are applicable to scalar random variables with uniform prior distributions. This bound was recently extended by Bell-Ephraim-Steinberg-Van Trees [4] to vectors of random variables with arbitrary prior distributions. The goal of this paper is to present an improvement to the vector version of [4], explore some properties of the bounds, and present further generalizations.

II. Improved Vector Bound

Assume that a vector of random variables θ with prior probability density function $p(\theta)$ is estimated from the observation vector $\mathbf{x}$. The estimation error is defined as $\epsilon = \hat{\theta} - \theta$ and we can lower bound $\mathbf{a}^T E\left\{\epsilon\epsilon^T\right\}\mathbf{a}$, where $\mathbf{a}$ is an arbitrary vector, by either of the two bounds:

$$(i)\quad \mathbf{a}^T E\left\{\epsilon\epsilon^T\right\}\mathbf{a} \geq \int_0^\infty \frac{\Delta}{2} \cdot V\left\{\int \left(p(\varphi)+p(\varphi+\delta)\right) P_{\min}(\varphi,\varphi+\delta)d\varphi\right\} d\Delta \tag{1}$$

$$(ii)\quad \mathbf{a}^T E\left\{\epsilon\epsilon^T\right\}\mathbf{a} \geq \int_0^\infty \frac{\Delta}{2} \cdot V\left\{\int 2\min\left(p(\varphi),p(\varphi+\delta)\right) P^{el}_{\min}(\varphi,\varphi+\delta)d\varphi\right\} d\Delta \tag{2}$$

where the vector δ satisfies

$$\mathbf{a}^T\delta = \Delta, \tag{3}$$

$P_{\min}(\varphi,\varphi+\delta)$ is the minimum probability of error in the binary detection problem:

$$\begin{aligned} H_0:&\quad \theta=\varphi; \qquad \Pr(H_0) = \frac{p(\varphi)}{p(\varphi)+p(\varphi+\delta)} \\ H_1:&\quad \theta=\varphi+\delta; \qquad \Pr(H_1) = 1-\Pr(H_0), \end{aligned} \tag{4}$$

$P^{el}_{\min}(\varphi,\varphi+\delta)$ is the minimum probability of error in the same binary detection problem but with equally likely hypotheses, and $V\{\cdot\}$ is a valley-filling function. Since probability of error results are easier to derive and more plentiful for the equally likely problem, the second bound may be more useful computationally.

In applying the bounds, one has to choose $\mathbf{a}$ and δ. The choice for $\mathbf{a}$ is dictated by the particular parameter or linear combination of parameters being investigated. If a bound on the MSE of the i^{th} parameter is desired, then $\mathbf{a}$ must be the unit vector with a one in the i^{th} position.

The vector δ determines the position of the second hypothesis $\theta = \varphi+\delta$. It is constrained to lie in the hyperplane defined by (3). Generally, δ should be chosen so that the two hypotheses are as indistinguishable as possible by the optimum detector. In [4], δ was chosen to be

$$\delta = \frac{\Delta}{\|\mathbf{a}\|^2}\,\mathbf{a}. \tag{5}$$

This choice results in the hypotheses being separated by the smallest Euclidean distance. When $\mathbf{a}$ is chosen to produce a bound on the MSE of the i^{th} parameter, the resulting bound is equivalent to that obtained by conditioning the scalar bound [4] on the remaining $i-1$ parameters, and taking the expected value with respect to those parameters. Choosing δ according to (5) does not always lead to the tightest bound because hypotheses separated by the smallest Euclidean distance are not necessarily the most indistinguishable. A higher $P_{\min}(\varphi,\varphi+\delta)$ can be achieved by choosing

$$\delta = \frac{\Delta}{\|\mathbf{a}\|^2}\,\mathbf{a} + \mathbf{b} \tag{6}$$

where $\mathbf{b}$ is not a function of φ, and is orthogonal to $\mathbf{a}$, thus (3) is satisfied. With this choice, the bound on the MSE of the i^{th} parameter cannot be reduced to the expected value of the conditional scalar bound.

III. Further Extensions

Other results concerning the Ziv-Zakai bound are as follows. First, a tighter bound which uses the probability of error in an M-ary hypothesis testing problem is derived.

Second, both the binary and M-ary bounds are equal to the minimum MSE when the posterior density of $\mathbf{a}^T\theta$ given the observations, $p(\mathbf{a}^T\theta|\mathbf{x})$, is symmetric and unimodal. Furthermore, in the limit of no data, the bound converges to the true a priori variance when the prior density of $\mathbf{a}^T\theta$ is symmetric and unimodal.

Third, for problems in which some of the parameters may be considered random variables with prior probability density functions, but some are considered unknown, deterministic quantities, the bound can be extended to a hybrid version combining the derivation leading to (1) and (2) with a derivation similar to that in [1] for non-random parameters.

Fourth, the bound can be extended to any non-decreasing cost function of $|\mathbf{a}^T\epsilon|$ in a straightforward manner.

References

[1] J. Ziv and M. Zakai, "Some Lower Bounds on Signal Parameter Estimation", *IEEE Trans. Information Theory*, vol. IT-15, no. 3, pp. 386-391, May 1969.

[2] D. Chazan, M. Zakai, and J. Ziv, "Improved Lower Bounds on Signal Parameter Estimation", *IEEE Trans. Information Theory*, pp. 90-93, January 1975.

[3] S. Bellini and G. Tartara, "Bounds on Error in Signal Parameter Estimation", *IEEE Trans. Comm.*, vol. 22, pp. 340-342, March 1974.

[4] K. L. Bell, Y. Ephraim, Y. Steinberg, and H. L. Van Trees, "Improved Bellini-Tartara Lower Bound for Parameter Estimation", in *Proceedings of 1994 International Symposium on Information Theory*, (Trondheim, Norway), June 1994.

Source Coding with a Reversible Memory-Binding Probability Density Transformation

Bryan G. Talbot* and Lisa M. Talbot†

* The Analytic Sciences Corp., Reston, VA

† 11031 Barton Hill Ct., Reston, VA 22091

Abstract — **We present a memory-binding density transformation as a means of improving performance of entropy coders acting on memoried sources.**

I. Introduction

Reversible coders are often called upon to operate on sources with memory. Though Shannon's work suggests that coding performance may be enhanced by encapsulating memory information in an M-dimensional pdf, in many situations this approach is impractical. Thus, coders are often forced to view the source as memoryless and attempt to encode near the entropy of a high-entropy single-symbol pdf, rather than the desired lower-entropy multi-symbol memoried pdf. We propose a reversible memory-binding transform (MBT) alternative which improves performance by binding memory information from the multi-symbol pdf into the single-symbol pdf to be processed by the coder.

II. Density Transformation Algorithm

Assume a source sequence $\{x_i\}$, $x_i \in A$, where $A = \{\alpha_1, \ldots, \alpha_N\}$ is an alphabet of N symbols. Memory information associated with x_i is represented by the vector $\mathcal{X}_i = \{x_{i-M}, \ldots, x_{i-1}\}$. For each x_i there exists a mapping ϕ_i, a permutation of A that produces $B_i = \phi_i A = \{\beta_1, \ldots, \beta_N\}$. This permutation is a function of $\mathcal{X}_i$ that encapsulates memory of the M symbols previous to x_i and may be represented by either a rule or list. In either case, ϕ_i has the property that the *a priori* probability, $P(x_i = \beta_1 \mid \mathcal{X}_i)$, is maximum and $P(x_i = \beta_n \mid \mathcal{X}_i) \geq P(x_i = \beta_{n+1} \mid \mathcal{X}_i)$ for $n = 1, \ldots, N-1$. The density transform is defined as $f : x_i \rightarrow y_i$ where $x_i, y_i \in A$, so that $y_i = f(x_i) = \sum_{n=1}^{N} \alpha_n \delta(x_i - \beta_n)$. The inverse transform is given by $x_i = f^{-1}(y_i) = \sum_{n=1}^{N} \beta_n \delta(y_i - \alpha_n)$. The probability density functions associated with x_i and y_i are given by $p(x)$ and $p(y)$ respectively.

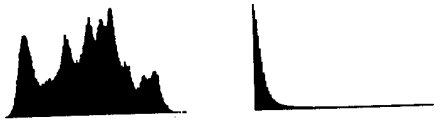
III. Transformation Characteristics

Viewed in one sense, an MBT is a generalization of both differential and modulo-PCM coding. In another sense, it is conceptually an alternative projection mechanism for producing a low-entropy low-dimensionality pdf from a low-entropy high-dimensionality pdf with memory. In a third sense, from an encoder perspective, it may also be viewed as a transformation from $p(x)$ to $p(y)$ which binds memory information to the symbols forming the domain y. From any perspective, an MBT, appropriately inserted between a source and coder, is a mechanism for increasing coder performance by reducing the entropy of the source pdf. For many sources of interest, such as imagery, $E\{p(y)\} < E\{p(x)\}$ where $E\{\bullet\}$ is the entropy function.

The fact that the MBT effectively reduces the entropy of a memoried source separately from the choice of encoder is significant from a theoretical perspective because it separates the coding process into distinct entropy reduction and coding stages rather than combining both operations into one step. Thus, the encoding process as a whole becomes more general and source-independent. The MBT also has the characteristics of transforming source densities of interest, which may be compound multi-modal distributions, into simple structured densities with a predictable parametric shape similar to the gamma distribution. This is useful from both an information theoretic and statistical perspective because it provides an effective interface between real world data sources and information theoretic models based on parametric distributions. In this case, coding models based on statistically determined gamma-type distributions may be directly exported to a variety of real-world sources via an MBT. The transform introduced here is similar to modulo coding schemes of [1] in that it does not increase the size of source alphabet supplied to the encoder. This contrasts with differential coding which can potentially increase the alphabet size by a factor of two.

IV. Experimental Results

The MBT is applicable to both traditional and non-traditional sources. Figure 1 demonstrates the application of the MBT to the ubiquitous Lena image for $M = 1$. Tests with MBT-AH (Adaptive Huffman) and MBT-LZW (Lempel-Ziv-Welch) coder pairs showed significant performance gains over those by either AH or LZW alone. We have also applied an MBT-AH coding pair to the indices output by a vector quantizer in conjunction with memory knowledge provided by the codebook. We have found that MBT-AH boosted VQ compression performance by a factor of nearly 1.5 in comparison to the factor of 1.1 for AH alone.



(a) (b)

Fig. 1: Reversible Transformation of Lena Image pdf: (a) $p(x)$, entropy = 7.44bpp (b) $p(y)$, entropy = 5.06bpp

V. Conclusions

We have presented a reversible memory-binding transform algorithm. The transform, inserted as a separate stage between the source and coder, serves to increase performance separately from choice of encoder. This transform shows much promise for use in the fields of information theory, coding, and statistics.

References

[1] K. Sayood and S. Na. Recursively Indexed Quantization of Memoryless Sources. *IEEE Transactions on Information Theory*, 38(5):1602–1609, 1992.

Projection Pursuit Autoregression and Projection Pursuit Moving Average

Zheng Tian
Applied Math. Dept., Northwestern Polytechnical Univ.,
710072 Xi' an Shaanxi, China

Abstract — **Projection pursuit autoregression and projection pursuit moving average with multivariate polynomials as ridge functions in both cases are proposed in this paper (To write the methods in simplifed forms MPPAR and MPPMA, respectively). The L_2-convergence of such the methods is proved. This paper also proposes two new algorithms for MPPAR and MPPMA. By using the methods, we establish the mathematical models about the Wolfer sunspot data and Canadian lynx data.**

I. Introduction

The results presented here were motivated by the research concerning both non-linear non-parametric time series analysis and projection pursuit regression.

II. Projection Pursuit Autoregression and Its L_2-Convergence

The process $x_t, t = 0, \pm 1, \pm 2, \cdots$ is said to be a non-linear autoregression of order k (NLAR(k)) process if x_t is stationary and there exists a function $f: R^k \to R$ such that $x_t = f(x_{t-1}, \cdots, x_{t-k}) + z_t, t \in Z$, where $z_t \sim \text{WN}(0, \sigma^2)$.
The form of projection pursuit autoregression model can be expressed as

$$x_t = \sum_{j=1}^{m} g_j(\alpha_j^T x) + z_t, t \in Z,$$

where $\alpha_j^T x$ denotes a one-dimensional projection of the vector $x^T = (x_{t-1}, \cdots, x_{t-k})$, α_j is projective direction, $\alpha_j^T = (\alpha_{j_1}, \cdots, \alpha_{j_k})$ and is called as ridge function and $z_t \sim \text{WN}(0, \sigma^2)$.

For the L_2-convergence of MPPAR we have the following Theorem:

Theorem. Let x_t be a non-linear autoregressive time series NLAR(k), and $\int f^2(x)dP < \infty$, then there exist α_j and $g_j(u)$ such that as $m \to \infty$,

$$\int [f(x) - \sum_{j=1}^{m} g_j(\alpha_j^T x)]^2 I_S(x) dP \to 0,$$

where $g_j(u)$ is a polynomial, $S = \{(x_1, \cdots, x_k): -c_1 \leqslant x_1 \leqslant c_1, \cdots, -c_k \leqslant x_k \leqslant c_k\}$, where $c_j, j = 1, \cdots, k$, are large enough positive numbers.

III. Projection Pursuit Moving Average and Its L_2-Convergence

Let $x_t, t \in Z$ be a non-linear moving average of order l (NLMA(l)) process, $x_t = h(z_{t-1}, \cdots, z_{t-l}) + z_t, t \in Z$, where $z_t \sim \text{WN}(0, \sigma^2)$, the function $h: R^l \to R$. It is obvious that MPPMA has the L_2-convergence as following corollary shows:

Corollary. Let x_t be a non-linear moving average time series NLMA(l), and $\int h^2(z)dP < \infty$, then there exist α_j and $g_j(u)$ such that as $m \to \infty$,

$$\int [h(z) - \sum_{j=1}^{m} g_j(\alpha_j^T z)]^2 I_S(z)\, dP \to 0,$$

where $g_j(u)$ is a polynomial and $S = \{(z_1, \cdots, z_l): -c_1 \leqslant z_1 \leqslant c_1, \cdots, -c_l \leqslant z_l \leqslant c_l\}$, where $c_j, j = 1, \cdots, l$ are large enough positive numbers. $z^T = (z_{t-1}, \cdots, z_{t-l})$.

IV. The Algorithms of MPPAR and MPPMA

We propose the following MPPAR algorithm:

step 1. First analysing the data and drawing the scatterplot, we use the Durbin-Levinson method and AIC criterion to identify the order k of the fitted model and estimate the variance of white noise.

step 2. We select the suitable m and the power-number of polynomials $g_j(.), j = 1, \cdots, m$ then minimizing the objective function

$$v(\alpha, g) = \sum_{i=1}^{N} [x_i - \sum_{j=1}^{m} g_j(\alpha_j^T x)]^2,$$

where $x^T = (x_{i-1}, \cdots, x_{i-k})$, x_i and x are observations, x_i is R valued and x is R^k valued $i = 1, \cdots, N$. We can establish the preliminary model, $x_t^{(1)} = \sum_{j=1}^{m} g_j(\alpha_j^T x) + z_t$.

step 3. Examining the residuals, we find out whether the residuals have the appearance of a realization of white noise.

step 4. Repeat step 2 and 3, until the residuals can be the appearance of a realization of white noise.

The algorithm of MPPMA is similar to the algorithm of MPPAR, except for the step 1.

V. The Applications

By using the new methods, we establish the models for the wolfer sunspot data and Canadian lynx data, respectively.

The mathematical model of the Wolfer sunspot data (1700-1970)

$x_t = -0.01431 + 0.29984\,\alpha_1^T x + 0.00555\alpha_1^T x - 0.00602 (\alpha_2^T x)^2 + z_t$, where $x^T = (x_{t-1}, x_{t-2})$, $\alpha_1^T = (0.88869, -0.45851)$, $\alpha_2^T = (-0.65241, 0.75786)$. $z_t \sim \text{WN}(0, 0.009)$.

The results show that the residuals of the model are less than 5%.

The mathematical model for Canadian lynx data (1821-1934).

$x_t = -0.02586 - 0.0415\,\alpha_1^T x - 0.57932\alpha_2^T x - 0.02586 (\alpha_2^T x)^2$, where $x^T = (x_{t-1}, x_{t-2})$, $\alpha_1^T = (0.87228, -0.48901)$, $\alpha_2^T = (0.09684, -0.99530)$. $z_t \sim \text{WN}(0, 0.023)$.

The results show that the residuals of the model are less than 2.5%.

Acknowledgements

This research work was partly supported by the Post-doc scholarship of University of Dortmund, Germany.

References

[1] J. H. Friedman, and W. Stuetzle, Projection pursuit regression. *J. Am. Statist. Assoc.* vol. 76 pp. 817-23, 1981.

ROOT-N CONSISTENT ESTIMATORS OF ENTROPY FOR DENSITIES WITH UNBOUNDED SUPPORT

A.B. Tsybakov
LSTA, Université Pierre et Marie Curie, Paris - France

E.C. van der Meulen
Department of Mathematics, Katholieke Universiteit Leuven, Leuven - Belgium

Abstract - **We consider a truncated version of the entropy estimator proposed in [1] and prove the mean square $\sqrt{n}$-consistency of this estimator for a class of densities with unbounded support, including the Gaussian density.**

SUMMARY

Let $X_1, ..., X_n$ be a sample of i.i.d. random variables with common density $f(x), x \in \mathbb{R}$. We consider the problem of estimating the unknown entropy

$$H(f) = -\int f(x) \ln f(x) dx.$$

This problem has various applications in hypothesis testing and information theory. There exist two main approaches to the construction of entropy estimators. The first approach consists of substitution of $f(x)$ in $H(f)$ by a suitable nonparametric density estimator. The second approach is based on spacings. Let $X_{n,1} \leq X_{n,2} \leq ... \leq X_{n,n}$ be the order statistics of $X_1, ..., X_n$. The estimator

$$\hat{H}_{m,n} = \frac{1}{n} \sum_{i=1}^{n-m} \ln(\frac{n}{m}(X_{n,i+m} - X_{n,i})),$$

where m is a positive integer less than n, was introduced in [2]. Its asymptotic properties as $n \to \infty$ have been studied by several authors under various assumptions on f. Here we study an entropy estimator which is somewhat different from $\hat{H}_{m,n}$ and is defined by

$$H_n = \frac{1}{n} \sum_{i=1}^{n} \ln\{2\rho_i \gamma(n-1)\}$$

where $\rho_i = \min\{a_n, \min_{j \neq i} |X_i - X_j|\}$, $a_n \to 0$ is a sequence of positive numbers, $\gamma = \exp\{C_E\}$, and C_E is Euler's constant. H_n is a truncated and modified version of the estimate introduced in [1]. In Theorem 1 we prove that the bias of H_n is of order $O(\frac{1}{\sqrt{n}}), n \to \infty$. In Theorem 2 we show that the variance of H_n is of order $O(\frac{1}{n})$. Our results hold for densities f with unbounded support and exponentially decreasing tails, such as the Gaussian density.
Consider the following assumptions :

(A_0) $\int f(x)|\ln f(x)|dx < \infty$.

(A_1) f is twice continuously differentiable and strictly positive on $\mathbb{R}$.

(A_2) $\int f(x) \exp(-bf(x))dx \leq Cb^{-1}$

where C is a finite positive constant.

Theorem 1. Assume $(A_0) - (A_2)$. Then, as $n \to \infty$,

$$E(H_n) - H(f) = O(\frac{1}{\sqrt{n}}).$$

Next consider the conditions :

(B_1) f is Lipschitz continuous and strictly positive on $\mathbb{R}$.

(B_2) There exists $a > 0$ such that for $j = 1, 2, 3$

$$\int f(x)(\frac{\sup_{|z-x| \leq a} f(z)}{f(x)})^j dx < \infty$$

$$\int f(x)(\frac{\sup_{|z-x| \leq a} f(z)}{f(x)})^j \ln^2 f(x) dx < \infty.$$

Theorem 2. Assume $(A_0), (B_1), (B_2)$. Then, as $n \to \infty$,

$$E(H_n - E(H_n))^2 = O(\frac{1}{n}).$$

Corollary. Assume $(A_0) - (A_2), B_2$. Then H_n is $\sqrt{n}$-consistent in the mean square, i.e., as $n \to \infty$

$$E(\sqrt{n}(H_n - H(f)))^2 = O(1).$$

REFERENCES

[1] L.F. Kozachenko, N.N. Leonenko, "Sample estimate of the entropy of a random vector", *Problems Inform. Transmission*, 23, 95-101, 1987.

[2] O. Vasicek, "A test for normality based on entropy", *J. Roy. Statist. Soc. Ser. B*, 38, 54-59, 1976.

On the Theory and Application of Universal Classification to Signal Detection

N. Warke and G. C. Orsak
Dept. of Electrical and Comp. Engineering
George Mason University
Fairfax, VA 22030-4444
E-mail: warkenc@bass.gmu.edu E-mail: gorsak@tejas.gmu.edu

Abstract– Herein we apply methods of Universal Classification to the problem of classifying one of M deterministic signals in the presence of dependent non-Gaussian noise.

Definition of Problem: The M-ary Signal Classification problem considered is defined as follows:

$$H_i \ : \ \mathbf{X}^n \sim \text{Noise} + \mathbf{S}(\theta_i) \quad i = 1, 2...M$$

where the test vector $\mathbf{X}^n$ is of length n. The $M+1^{th}$ hypothesis corresponds to the case in which $\mathbf{X}^n$ arises from none of the above M hypothesis. In addition, we shall assume that the noise arises from an unknown K^{th} order Markov source and that hypothesis H_1 corresponds to "no signal present," that is $\mathbf{S}(\theta_1) = 0$ for all n.

In this work we develop a classification scheme which is independent of the true statistical model of the environment and still achieves many of the desirable properties of the globally optimal detector.

Proposed Classifier: In the absence of a statistical model for the noise, we will assume the existence of a length N training vector $\mathbf{t}^N$ from the noise source. We propose the following classifier based on the work of Ziv and Gutman in [1,2]:

$$h_q(\mathbf{X}^n, \mathbf{t}^N, \theta_i, \lambda) = d_{KL}(P_{(\mathbf{X}^n - \mathbf{S}(\theta_i))_q}, P_{((\mathbf{X}^n - \mathbf{S}(\theta_i)), \mathbf{t}^N)_q}) + \frac{N}{n} d_{KL}(P_{\mathbf{t}_q^N}, P_{((\mathbf{X}^n - \mathbf{S}(\theta_i)), \mathbf{t}^N)_q}) - \lambda \quad , i = 1, 2, ..., M$$

where $(.)_q$ denotes the quantization of the continuous alphabet source, $d_{KL}(P_{X_q^n}, P_{Y_q^m}) \stackrel{def}{=} \sum P_{X_q^n} \log\{\frac{P_{X_q^n}}{P_{Y_q^m}}\}$, is the Kullback -Leibler distance between the types $P_{X_q^n}$ and $P_{Y_q^m}$ of the quantized data and λ is a positive constant chosen to satisfy some design criterion. The decision regions $\{\Lambda_1, \Lambda_2, ..., \Lambda_M\}$ corresponding to the M hypotheses $H_1, H_2, ..., H_M$ are defined as follows: The region Λ_1 is defined as the set of all sequences $(\mathbf{X}^n, \mathbf{t}^N)$ for which $h_q(\mathbf{X}^n, \mathbf{t}^N, \theta_i, \lambda) \geq 0$ for all $i = 2, 3...M$. The region Λ_j for $j = 2, 3...M$ is defined as the set of all sequences $(\mathbf{X}^n, \mathbf{t}^N)$ for which $h_q(\mathbf{X}^n, \mathbf{t}^N, \theta_i, \lambda) \geq 0$ for all $i = 1, 2...M, i \neq j$ and $h_q(\mathbf{X}^n, \mathbf{t}^N, \theta_j, \lambda) < 0$ and the rejection region $\Lambda_R = (\bigcup_{i=1}^M \Lambda_i)^c$.

Summary of Theoretical Results:

1) The asymptotic probability of error under each hypothesis decays exponentially fast at a rate λ as the length of the test vector $\mathbf{X}^n$ grows without bound,

$$\lim_{n\to\infty} \frac{1}{n} \log P_\Lambda(e/H_i) \leq -\lambda, i = 1, ...M$$

regardless of the length of the training vector.

2) The asymptotic probability of detection under each hypothesis tends to 1 as the length of the test vector grows without bound provided that $\lim_{n\to\infty} \frac{N}{n} > 0$ and $0 < \lambda \leq \lambda_o$,

$$\lim_{n\to\infty} P_\Lambda(\Lambda_i/H_i) = 1, i = 1, ...M$$

for an appropriately chosen constant λ_o.

3) The probability of rejection under each hypothesis for iid noise sources falls off exponentially fast as the length of the test vector grows without bound subject to the above constraints on n, N and λ,

$$\lim_{n\to\infty} \frac{1}{n} \log P_\Lambda(\Lambda_R/H_j) \leq -\lambda_j, j = 1, 2, ...M$$

where $\lambda_j > 0$, $j = 1, 2, ...M$.

References:

[1] J. Ziv, "On Classification with Empirically Observed Statistics and Universal Data Compression", *IEEE Trans. Inform. Theory*, vol. 34, pp. 278-286, Mar 1988.

[2] M. Gutman, "Asymptotically Optimal Classification for Multiple Tests with Empirically Observed Statistics", *IEEE Trans. Inform. Theory*, vol. 35, pp. 401-408, Mar 1989.

A Matrix Form of the Brunn-Minkowski Inequality and Geometric Rates

Ram Zamir and Meir Feder

330 E&TC Building, Cornell University, Ithaca, NY 14853 . e-mail: zamir@ee.cornell.edu

Dept. of Electrical Eng. - Systems, Tel-Aviv University, Tel-Aviv 69978 Israel . e-mail: meir@eng.tau.ac.il

Abstract — **A matrix form of the Brunn-Minkowski Inequality is derived, which may be applied in calculating the uncoded bit rate of lattice quantization and modulation schemes.**

I. Introduction

In many practical coding schemes, the bit allocation to the symbols, at least in the intermediate phases of the coding process, is done according to *geometric* consideration rather than to probabilistic ones. For instance, the number of effective code words of a lattice quantizer is determined by the shape and the volume of its granular region, and by the shape and the volume of its Voronoi cell. Similarly, the size of a lattice type constellation of a coded modulation system is determined by the shape and the volume of the symbols' decision cells, and the shape and the volume of the space of allowable signals (the latter is determined, e.g., by peak power, peak spectrum or peak amplitude constraints).

Hence, the bit rate at the quantizer output, or in the modulator input, is generally higher than the overall coding rate of the system, and it may estimated by the *geometric rate*

$$R_G \sim \log\left(\frac{\mu(\mathcal{A}_X+\mathcal{A}_N)^{1/d}}{\mu(\mathcal{A}_N)^{1/d}}\right) , \qquad (1)$$

where d is the dimension of the space (of source or channel signals), $\mathcal{A}_X$ is the region of input signals, $\mathcal{A}_N$ is the basic cell (of the quantizer or the constellation), $\mu(\mathcal{A}) = \int_{x\in\mathcal{A}} dx$ is the (d-dimensional) volume of the region $\mathcal{A}$, and $\mathcal{A}_X + \mathcal{A}_N = \{x+y : x\in\mathcal{A}_X, y\in\mathcal{A}_N\}$ is the *Minkowski sum* of $\mathcal{A}_X$ and $\mathcal{A}_N$. This sum may be interpreted as the geometric convolution of the two regions.

As in the problem of estimating the information rate in an additive noise channel, the geometric rate R_G is also not calculated easily, but it may be estimated by means of lower and upper bounds. For example, the volume of the Minkowski sum in (1) may be lower bounded via the *Brunn-Minkowski Inequality* (BMI), [1],

$$\mu(\mathcal{A}_X+\mathcal{A}_N)^{1/d} \geq \mu(\mathcal{A}_X)^{1/d} + \mu(\mathcal{A}_N)^{1/d} . \qquad (2)$$

Equality in (2) holds if the two regions are convex and proportional, e.g., if they are balls or cubes (with parallel edges). For $d=1$, this condition is reduced to the simple case where $\mathcal{A}_X$ and $\mathcal{A}_N$ are intervals (and not, e.g., a union of intervals).

The BMI is dual in some sense to the Entropy-Power Inequality (EPI), which lower bounds the entropy-power of the sum of independent random variables. In [2], a matrix form for the EPI was derived, leading to tight lower bounds on the capacity of an additive noise channel with memory or with intersymbol interference. In parallel to that, we derive in this work a matrix form for the BMI, which enables to give a tight estimate for R_G in cases where linear transformations are incorporated with coding ("shaping"), or when spectral constraints upon the signals are given.

II. Linear Transformation of Shapes

We first introduce the matrix form of the Minkowski sum. Let $\underline{\mathcal{A}}^t = (\mathcal{A}_1 \ldots \mathcal{A}_n)$ be a (row) vector, whose n components are d-dimensional shapes. We define a linear transformation of $\mathcal{A}_1 \ldots \mathcal{A}_n$ as

$$T\underline{\mathcal{A}} = \{T\underline{x} : x_i \in \mathcal{A}_i \text{ for } i = 1\ldots n\} , \qquad (3)$$

where T is an $m \times n$ matrix. In particular, $t\mathcal{A}$ means scaling the coordinates of $\mathcal{A}$ by the scalar t. Note that $T\underline{\mathcal{A}}$ is an md-dimensional shape. Denote the volumes of the shapes by $\mu(\mathcal{A}_i) = \mu_i, i = 1\ldots n$. Following simple laws of integration, the md-dimensional volume of $T\underline{\mathcal{A}}$, in the particular case $m = n$, is $\mu(T\underline{\mathcal{A}}) = |T|^d \cdot \mu(\underline{\mathcal{A}}) = |T|^d \cdot \prod_{i=1}^n \mu_i$, where $|\cdot|$ denotes the absolute value of the determinant. For the general case, we suggest the following matrix generalization of the BMI:

Theorem 1 (Matrix-BMI): *Let $\underline{\widetilde{\mathcal{A}}}^t = (\widetilde{\mathcal{A}}_1 \ldots \widetilde{\mathcal{A}}_n)$ be a vector of d-dimensional* cubes *whose edges parallel the axes, and whose volumes are the same as of $\mathcal{A}_1 \ldots \mathcal{A}_n$, i.e., $\mu(\widetilde{\mathcal{A}}_i) = \mu_i, i = 1\ldots n$. Then*

$$\mu(T\underline{\mathcal{A}})^{1/d} \geq \mu\left(T\underline{\widetilde{\mathcal{A}}}\right)^{1/d} = \sum_{i=1}^{\binom{n}{m}} |\widetilde{T}_i| \qquad (4)$$

where $\widetilde{T} = T \cdot L$, L is an $n \times n$ diagonal matrix whose diagonal elements are $\mu_1^{1/d} \ldots \mu_n^{1/d}$ (the edges' lengths of the cubes $\widetilde{\mathcal{A}}_1 \ldots \widetilde{\mathcal{A}}_n$), and $\left\{\widetilde{T}_i, i = 1\ldots\binom{n}{m}\right\}$ is the set of all possible $m \times m$ sub-matrices of $\widetilde{T}$, obtained by choosing m out of the n columns of $\widetilde{T}$.

For $m = 1$, (4) reduces to $\mu\left(\sum_{i=1}^n t_i\mathcal{A}_i\right)^{1/d} \geq \sum_{i=1}^n |t_i|\mu_i^{1/d}$, i.e., to the regular BMI (2). Equality in (4) holds in each one (or in a mixture) of the following cases: if $\mathcal{A}_1 \ldots \mathcal{A}_n$ are cubes whose faces parallel each other; if (after removing the all zero columns of $\widetilde{T}$, if any) $m = n$; or if $\widetilde{T}$ does not have a full row rank, where then $\mu(T\underline{\mathcal{A}}) = 0$. Theorem 1 is proved via a double induction over the dimensions of T, using a conditional form of the BMI, analogously to the proof of the matrix-EPI in [2].

Acknowledgements

This research was supported in part by the Wolfson Research Awards administered by the Israel Academy of Science and Humanities. The authors also wishes to thank helpful comments by Shlomo Shamai and Simon Litsyn.

References

[1] A. Dembo, T.M.Cover, and J.A.Thomas. Information theoretic inequalities. *IEEE Trans. Information Theory*, IT-37:1501–1518, Nov. 1991.

[2] R. Zamir and M. Feder. A generalization of the Entropy Power Inequality with applications. *IEEE Trans. Information Theory*, IT-39:1723–1727, Sept. 1993.

POSTER SESSION II

Constructing Wavelets from Desired Signal Functions

Major Joseph O. Chapa, USAF
Mysore Raghuveer

Center for Imaging Science, Rochester Institute of Technology, Rochester, NY 14623, USA
Dept. of Electrical Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA

Abstract — **A limitation to wavelet design is the inability to construct orthonormal wavelets that match or are "tuned" to a desired signal. This paper develops a technique for constructing an orthonormal wavelet that is optimized in the least squares sense, and whose associated scaling function generates an orthonormal multiresolution analysis (OMRA).**

I. Introduction

Most applications of orthonormal multiresolution analyses (OMRA) use either Daubechies', Meyer's, or Lemarie's wavelets [1, 2, 3]. However, it would be best if the wavelet matched the signal of interest. This paper presents a technique for generating an OMRA with a wavelet that is matched in the least squares sense to a signal of interest by first developing a method for constructing the scaling function from the wavelet and second, giving the conditions on the wavelet that guarantee an OMRA.

II. Multiresolution Decomposition

Mallat [1] showed that the discrete wavelet transform can be used to generate an orthonormal multiresolution decomposition of a discrete signal consisting of a series of detail functions and a residual low resolution approximation of the original signal. The decomposition is done by convolving the original sequence with a pair of quadrature mirror filters, h (low pass) and g (high pass). In order to perfectly reconstruct the original signal from the detail functions and the residual approximation, the following must be true of the Fourier spectrum magnitudes of h and g.

$$|H(\omega)|^2 + |G(\omega)|^2 = 1 \tag{1}$$

Cancellation of any aliasing is guaranteed by setting $g_k = (-1)^k h_{1-k}$. The filters, h and g, are related to the mother wavelet, $\psi(x)$, and the scaling function, $\phi(x)$, by their 2-scale relations [2], $\psi(x) = 2\sum_k g_k \phi(2x-k)$ and $\phi(x) = 2\sum_k h_k \phi(2x-k)$, or in the frequency domain by

$$\Psi(\omega) = G(\frac{\omega}{2})\Phi(\frac{\omega}{2}) \qquad \Phi(\omega) = H(\frac{\omega}{2})\Phi(\frac{\omega}{2}) \tag{2}$$

III. Constructing Φ from Ψ

A recursive equation for finding $\Phi(\omega)$ from $\Psi(\omega)$ can be found by taking the magnitude squared of the equations in (2), adding them and substituting equation (1) giving

$$|\Phi(\omega)|^2 = |\Phi(2\omega)|^2 + |\Psi(2\omega)|^2 \tag{3}$$

Substituting $\omega = \pi k$, then $\omega = \pi k/2$ and so on, leads to the following closed form solution.

$$\left|\Phi\left(\frac{\pi k}{2^\ell}\right)\right|^2 = \sum_{n=0}^{\ell} \left|\Psi\left(\frac{2\pi k}{2^n}\right)\right|^2 \text{ for } k \neq 0 \tag{4}$$

and as in Mallat[1], $\Phi(0) = 1$. So, given any known wavelet, its corresponding scaling function can be found directly from equation (4).

IV. Guaranteeing Orthonormality

Given that $g_k = (-1)^k h_{1-k}$ and $\Phi(0) = 1$, the multiresolution generated by $\phi(x)$ and related to $\psi(x)$ is orthonormal[1, 2] if $<\phi(x), \phi(x-k)> = \delta(k)$, or in the frequency domain, $\sum_{m=-\infty}^{\infty} |\Phi(\omega + 2\pi m)|^2 = 1$. Applying this condition to equation (4) and letting $\Delta\omega = \pi/2^\ell$ gives the following condition on $\Psi(\omega)$ that will guarantee an orthonormal multiresolution analysis.

$$\sum_{m=-\infty}^{\infty} \sum_{n=0}^{\ell} \left|\Psi\left(2^{n+1}\left(\frac{2\pi k}{2^{\ell+1}} + 2\pi m\right)\right)\right|^2 = 1 \tag{5}$$

Any wavelet that satisfies condition (5) can be used in equation (4) to generate an orthonormal scaling function that, in turn, generates an orthonormal multiresolution analysis.

V. Finding Matched Wavelets

Assume in equation (5) that $\Psi(\omega)$ is bandlimited to $\pi \cdot K_L < |\omega| < \pi \cdot K_U$. Then for each value of $\ell = 0, 1, \ldots, N$ where $\Delta\omega = \pi/2^N$ is the sample spacing chosen for $\Phi(k\Delta\omega)$, a set of M equality constraints on $\Psi(k\Delta\omega)$ can be derived with the following form

$$\sum_{i=1}^{M} \alpha_{ik} \left|\Psi\left(\frac{k\pi}{2^N}\right)\right|^2 - 1 = 0 \tag{6}$$

where $\alpha_{ik} = \{0, 1\}$. Given $W(k\Delta\omega)$ as the desired signal spectrum, the equality constraints in (6) along with the inequality constraints, $0 \leq |\Psi(k\pi/2^N)|^2 \leq 1$ can be solved using nonlinear programming techniques where the objective function $f = \sum_k (|W(k\Delta\omega)| - |\Psi(k\Delta\omega)|)^2$ is minimized. The result is a wavelet spectrum that satisfies the conditions for orthonormality and is matched to the desired spectrum, $W(k\Delta\omega)$. Since the resultant wavelet spectrum is magnitude only, the wavelet is symmetrical.

References

[1] S. G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 11, no. 7, July 1989.

[2] C. K. Chui, "An Introduction to Wavelets," *Wavelet Analysis and Its Application, Vol I*, Academic Press, Inc., 1992.

[3] I. Daubechies, "Orthonormal Bases of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics*, v. 41, 1988.

Wavelet Transform based ECG Data Compression with Desired Reconstruction Signal Quality

Jie Chen†, Shuichi Itoh†† and Takeshi Hashimoto†

† Dept. E.E., †† Grad. Sch. IS. , Univ. of Electro-Communications, Chofu, Tokyo 182, Japan, E-mail: chen@ee.uec.ac.jp

Abstract — **This paper proposes a new coding strategy by which desired quality of reproduced signal can be guaranteed in the minimum cost of coding rate.**

I. Introduction

In [1], we introduced discrete orthonormal wavelet transform (DOWT) into the ECG data compression and obtained the results of compression ratios (CR) from 13.5 : 1 to 22.9 : 1 with the corresponding percent *rms* difference (PRD) between 5.5% and 13.3%. Although we have achieved a dramatic improvement on compression performance over traditional ECG compression schemes, both in our previously proposed ECG coding system and in other lossy compression schemes, there exists an unresolved problem that the desired quality of the reproduced signals is hardly guaranteed in the minimum cost of coding rates. For ECG compression, such a problem becomes extremely crucial because if the quality of the reproduced signals can not be guaranteed the compression will be useless. The current solution to this problem is to sacrifice the coding rates in order to maintain the quality of reconstructed signals. Obviously, this is not an efficient manner. In this paper, we propose a DOWT-based compression scheme by which desired PRD between the original ECG signal and the reproduced signal can be guaranteed in the minimum cost of coding rate.

II. Coding Strategy

The DOWT-based ECG coding system is a hierarchical structure composed of three parts, the DOWT unit, the quantizing unit and the entropy coding unit. By a J-layered such a coding system, an input discrete signal $\mathbf{a}_0$ is progressively decomposed into a set of sub-signals $\{\mathbf{a}_J, (\mathbf{d}_j)_{1\le j\le J}\}$, where $\mathbf{a}_J$ is the lowest frequency sub-signal and $\{(\mathbf{d}_j)_{1\le j\le J}\}$ are the differential details at different frequencies. These sub-signals are quantized, entropy-encoded and transmitted to the receiver. At receiving side, the quantized sub-signals, $\{\mathbf{a}'_J, (\mathbf{d}'_j)_{1\le j\le J}\}$ are used to reconstruct the original signal. Let ε_j^d, ε_J denote the quantization MSE's of $\mathbf{d}_j$ and $\mathbf{a}_J$, respectively, according to [2], the reconstruction MSE between $\mathbf{a}_0$ and its reproduction is given by

$$\gamma = \varepsilon_J + \sum_{j=1}^{J} \varepsilon_j^d. \tag{1}$$

Based on Eq. (1), the problem of guaranteeing a desired reconstruction MSE γ_0 in the minimum cost of coding rate can be formulated as

$$\textbf{minimize } \overline{h(\gamma_0)} = \frac{1}{2^J} h_J + \sum_{j=1}^{J} \frac{1}{2^j} h_j^d \tag{2}$$

$$\textbf{subject to } \gamma = \gamma_0, \tag{3}$$

where, $\overline{h(\gamma_0)}$ is the output, h_J and h_j^d are the entropies of $\mathbf{a}_J$, $\mathbf{d}_j$ after quantization, respectively. And γ is the reconstruction MSE between $\mathbf{a}_0$ and its reproduction given by Eq. (1). The following optimum solution can be obtained by using Lagrange multiplier

$$\varepsilon_j^d = \frac{1}{2^j}\gamma_0, \quad \text{and} \quad \varepsilon_J = \frac{1}{2^J}\gamma_0. \tag{4}$$

Therefore, as soon as the quantization MSE's determined by Eq. (4) are achieved, the desired reconstruction MSE γ_0 (or the corresponding PRD) can be obtained.

How to achieve the desired quantization MSE's determined by Eq. (4) is another key point to realize our coding strategy. In what follows, we propose an adaptive quantizer by which the desired quantization MSE is really achievable. For a uniform quantizer, the quantization MSE is given by $\varepsilon_i = K(\Delta_i)^2$, where K is a constant factor, Δ_i is the quantization step-size. It is easy to see that, by adjusting the step-size Δ_i, the desired quantization MSE can be achieved. More details, for an expected quantization MSE, say, ε_0, we can randomly choose an initial step-size Δ_0 to do the quantization, then an actual quantization MSE σ_0 is obtained. We compare it with the expected ε_0, if a given precision is not satisfied, replace Δ_0 with $\sqrt{\varepsilon_0/\sigma_0}\,\Delta_0$ and repeat the process until a satisfactory precision is reached. Experiments have shown that the convergence usually finished within about 3 iterations.

III. Experiments

We have tested our proposed coding scheme at different desired PRD's. The ECG data is taken from the MIT-BIH Arrhythmia Database Record 200. The experimental results are shown in Table 1. The reconstructed ECG were evaluated by cardiologists and it seems clinically acceptable even at the CR as high as 22.7 : 1.

IV. Conclusion

In this paper, we proposed a new coding strategy by which desired quality (PRD) of reproduced signal can be guaranteed in the minimum cost of coding rate. The idea was successfully introduced to the DOWT-based coding system for the ECG compression application.

References

[1] Chen J., Itoh S. and Hashimoto T.: "ECG Data Compression by Using Wavelet Transform", *IEICE Trans. on Information Systems*, vol. E76-D, no. 12, pp. 1454–1461, 1993.

[2] Chen J., Itoh S. and Hashimoto T.: "Scalar quantization noise analysis and optimal bit allocation for a wavelet pyramid image coding system", *IEICE Trans. on Fundamentals*, vol. E76-A, no. 9, pp. 1502–1514, 1993.

Desired PRD(%)	7	9	13
Actual PRD(%)	6.7	8.9	13.2
CR	12.4:1	16.0:1	22.7:1

Table 1: Summary of the compression performance

APPLICATION OF MARKOV MODEL IN MOBILE COMMUNICATION CHANNEL

Wang Duanyi and Hu Zhengming
P. O. Box 145, Inform. Eng. Dept., Beijing Univ. of Posts and Telecom.
Beijing 100088, P. R. China

Abstract -- An adaptive Markov model with three states for mobile communication channel is studied and simulated. The error sequence describing the long burst error characteristics of the model channel is generated on a computer based on the model. A test method using threshold technique is given to verify the accuracy of the adaptive channel model.

I. INTRODUCTION

Modelling mobile communication channel is a prerequisite to improve the channel error performance by error-error-control technique. Set up a not only universal but also accurate statistic model for mobile communication channel is of great practical interest. In this paper, we apply Markov model with three states to setting up an adaptive statistic model for mobile communication channel based on a number of field test curve. The model parameters, i.e., the elements of the channel transition probability matrix P, are expressed as the functions of the average carrier-to-noise ratio (C/N). This is because among the factors which dominated the burst error characteristics of mobile channel, C/N plays an important role. We accomplish following work:

II. SETTING UP THE ADAPTIVE CHANNEL MODEL

First, we use a simple partitioned Markov model with three states as the probability statistic model to be set up describing the long burst error characteristics in mobile communication channel. The parameters of the state transition probability matrix $P=\{p_{ij}\}(i,j=1,2,3)$ of the model and the burst interval length distribution $G(m)=A_1e^{\alpha 1m}+A_2e^{\alpha 2m}$ of the channel error sequence have following relations:

$p_{11}=e^{\alpha 1}$ $p_{22}=e^{\alpha 2}$ $p_{31}=A_1e^{\alpha 1}$ $p_{32}=A_2e^{\alpha 2}$

$p_{13}=1-p_{11}$, $p_{23}=1-p_{22}$ $p_{33}=1-p_{31}-p_{32}$.

Then, according to a group of field test curves under different values of C/N, which is measured in a typical and mobile propagation environment, we set up the adaptive Markov model with three states with its parameters which are the functions of C/N by the curve fitting with the method of nonlinear least square.

III. GENERATING THE ERROR SEQUENCES BASED ON THE MODEL SET UP

According to the parameters of the adaptive channel model above set up, we generate an error sequence $\{e_i\}$, which describes the long burst error characteristic of mobile channel under different values of C/N, on a computer by following method:

1) Set i=0, assume an initial state s_0 (s_0=1, 2 or 3);

2) Generate a pseudo-random number r_i evenly distributed in the interval [0,1] by the hybrid congruence method.

3) Determine the value of e_i under the current state, and judge the next state according to:

a) If $s_i=3$, $e_i=1$,

$$s_{i+1}=\begin{cases}1 & \text{when } r_i<p_{31}\\ 2 & \text{when } p_{31}\le r_i<p_{31}+p_{32}\\ 3 & \text{when } p_{31}+p_{32}\le r_i\end{cases}$$

b) If $s_i=j\neq 3$, $e_i=0$,

$$s_{i+1}=\begin{cases}j & \text{when } r_i<p_{jj}\\ 3 & \text{when } r_i\ge p_{jj}\end{cases}$$

4) i=i+1, return 2).

After generating the error sequence, we estimate its performance by computing its burst interval length distribution probability G(m) and average bit error rate P_b' for corresponding value of C/N.

IV. MODEL ACCURACY TEST AND CONCLUSIONS

To verify the accuracy of the adaptive channel model above set up, we present a test method using threshold technique and its fundamental principle is as follows:

Since the burst error characteristics of mobile communication channel is regarded as a Rayleigh distribution (here only consider the case with severe fading), we can verify the accuracy of the adaptive channel model above set up by checking, for each value of C/N, if the burst error length distribution in $\{e_i\}$ generated accords with a Rayleigh distribution.

We can first generate a random number with Rayleigh distribution (denoted by s_i) from a random number evenly distributed in the interval [0,1] (denoted by r_i) according to

$$s_i=u\sqrt{-2\ln r_i} \qquad (1)$$

there u is the mean value of Rayleigh distribution, and therefore generate a random sequence $\{s_i\}$ with Rayleigh distribution. Then we set up a threshold B as follows

$$B=u\sqrt{-2\ln(1-P_b')} \qquad (2)$$

where P_b is average bit error rate of a random sequence. The threshold values for various values of C/N are obtained by letting P_b be equal to the P_b' under corresponding value of C/N computed in the above performance estimation.

According to above threshold value for each value of C/N, we quantize $\{s_i\}$ into a (0,1) sequence with Rayleigh distribution by following threshold comparition method. i.e.,

$$\begin{cases}e_i=0, & \text{when } s_i<B\\ e_i=1, & \text{when } s_i\ge B\end{cases}$$

For the (0,1) sequence with Rayleigh distribution, we estimate its performance by computing its burst interval length distribution G(m)" and average bit error rate P_b". Then by comparing G(m)" and P_b" with G(m)' and P_b' respectively under the same value of C/N, we observe that both are very close for each value of C/N, so the adaptive channel model is accurate to a great extent. Besides, compared with the conventional general partitioned Markov model where there are a number of error states, our model is more practical since it is easy to compute. In one word, it is a feasible scheme for optimizing mobile communication channel model.

REFERENCES

[1] X. Tao and D. Yuan, "Adaptive Markov model with three states for mobile communication channel," The 2nd National Commun. Conf. for Youths, 1991

[2] D. Yuan and K. Wu, "Setting up of Markov model with three states for the mobile radio channel and generating of the error sequence on a computer," J. of Shandong Univ., Vol.23, No.4, Dec. 1988

Markov Chains for Modeling and Analyzing Digital Data Signals

Richard Eier

Inst. of Computer Technology, Tech.Univ. of Vienna, Gusshausstr. 27-29, A-1040 Vienna, Austria

Abstract — **Digital data transmission signals may be considered as some specific stochastic process controlled by a Markov chain. Briefly going into the presentation and evaluation of the power density spectra (PDS) of such processes, one of our major concerns deals with the computational effort. By some special grouping among the employed signal elements and a corresponding partitioning of the controlling transition matrix, the formula for the desired PDS can be simplified to an Euclidean vector norm expression. By means of several PDS graphs the relevance of such an analysis to evaluate or design real transmission systems may be appreciated.**

1. Markov Chain Models

The signals usually employed for digital data transmission can be considered as some running sequences of discrete, especially shaped signal elements ([1]). Coding or modulating devices will generally introduce some memory into the system, thus the next signal element to be sent may depend on one or more of the previously sent elements. Therefore the Markov chain theory ([2]) provides very effective tools for exploring such signals. (e.g. [3],[4]).

In detail we consider an isochronological stationary Markov process (pace width = 1·T) that is internally controlled by a homogenous Markov chain. For the chain itself we denote the transition matrix $\underline{P} = (p_{k,l})$ and the absolute state distribution $\underline{q} = (q_1..q_K)$, either one being independent of the observation time. To prevent tedious discussion here, we assume the Markov chain to be ergodic ([2]), i.e. $\underline{P}^{\infty} \Rightarrow (1..1)^T \cdot (q_1..q_K)$ does exist and all q_k are > 0.

For the external process realization we consider signal elements that are individually assigned to the internal states. For convenience we specify them here in the frequency domain and denote them as $S_k(\omega)$. In any case the equality of $S_k(\omega) = S_l(\omega)$ assigned to different states is conformable with our model.

Modeling the system comprises first of all the decision on the abstract Markov states, the real signal elements and the relationship between them. Eventually the codulator operation and the statistics of the source data must be introduced into the state transition matrix. In an advanced model one may think about grouping the external signal elements in special classes such that some partitioning of the transition matrix becomes obtainable which can finally lessen the computational effort considerably.

2. The Power Density Spectra of Markov Processes

Here we are interested in evaluating the power density spectra (PDS) of our transmission signals. According to the Markov chain theory a basic formula can be derived (e.g. [3],[4] et alt.) which then may be rewritten in an Hermitean form using $\underline{E} = Diag(1..1)$, $\underline{Q} = Diag(q_1 .q_K)$, $\underline{s} = (S_1(\omega)...S_K(\omega))^T$ and $z = \exp(-j\omega t)$.

$$PDS_{cont} = \underline{s}^{+} \cdot ((\underline{I} - z\underline{P})^{+})^{-1} \cdot (\underline{Q} - \underline{P}^T\underline{Q}\underline{P}) \cdot (\underline{I} - z\underline{P})^{-1} \cdot \underline{s} \,. \qquad (1)$$

As a first application of this formula, we look at the maximum entropy process which is typically equipped with equally distributed state probabilities $\underline{q} = 1/K \cdot (1..1)$ or $\underline{Q} = 1/K \cdot \underline{I}$, respectively, and the transition matrix $\underline{P} = (1..1)^T \cdot \underline{q}$. In this special case the symmetric factor in the center of the matrix product (1) reads in particularly as $K^{-½} \cdot (\underline{I} - (1..1)^T \cdot \underline{q}) \cdot K^{-½}$. Now it is important to appreciate that the last symmetric matrix is idempotent ([2]). Thus it finally turns out that the former equation (1) is obviously equivalent to the following Euclidean vector norm expression

$$PDS_{cont} = \| K^{-½} \cdot (\underline{I} - (1..1)^T \cdot \underline{q}) \cdot \underline{s} \|^2 \,. \qquad (2)$$

For the more general statistically independent processes where the Markov states are distributed as $\underline{q} = (q_1..q_K)$, the analog result is easily verifiable:

$$PDS_{cont} = \| Diag(\sqrt{q_1}..\sqrt{q_K}) \cdot (\underline{I} - (1..1)^T \cdot \underline{q}) \cdot \underline{s} \|^2 \,. \qquad (3)$$

Eventually we may explore L-value FSK schemes with continuous phase characteristics. It is most profitable to organize the signal elements in a 2 level hierarchy: At first we define L classes of elements in respect to their frequency parameter f_k, and then we identify all possible phase values which occur at the start of each element and we denote the necessary quantity of them by M. Therefore the total of needed Markov states adds up to K=M·L.

The proposed structuring of the signal elements immediately motivates a conforming partitioning of the transition matrix using block matrices ([5]). Thus for statistically independent source data, one can formally establish $\underline{P} = (\underline{C}_M^{-i1} .. \underline{C}_M^{-iL})^T \cdot (q_1 \cdot \underline{I}_M .. q_L \cdot \underline{I}_M)$, where $\underline{C}$ stands for a cyclic matrix, the exponents refer to the phase differences between the beginning and end of the various signal elements of frequency f_k, and the subscripts indicate the matrix' dimension. Using the notion of a Kronecker product [5] one can still rewrite $\underline{P} = (\underline{C}_M^{-i1} .. \underline{C}_M^{-iL})^T \cdot (q_1..q_L) \otimes \underline{I}_M$.

Following the procedure of the previous examples we finally arrive again at an Euclidean vector norm expression for the PDS.

$$PDS_{cont} = \left\| Diag(\sqrt{q_1}..\sqrt{q_L}) \otimes \underline{I}_M \left(\underline{I}_L - (\underline{C}_M^{-i_1} .. \underline{C}_M^{-i_L})^T (q_1..q_L) \right) \otimes \underline{I}_M (\underline{I} - z\underline{P})^{-1} \underline{s} \right\|^2 \qquad (4)$$

Although this expression could be further evaluated in general form we won't discuss this issue here in anymore detail.

3. Applications and Future Work

In the project "DIG-SPEC - Power Spectra of Digital Data Signals", the PDS formulas for several modulated carrier as well as for baseband coded signals were evaluated. A program package for computation using these formulas and displaying the graph on standard VDU is already available. This program system will support investigations of frequency characteristics and bandwidth requirements of existing or new codulation schemes.

In theory we will carry out further studies of Markov processes employing transition matrices which have block structure and are particularly capable for treatment by Kronecker products.

4. References

[1] J.W.Smith,"A Unified View of Synchronous Data Transmission System Design", Bell Syst.Tech.J.,vol.47, pp.273-300, 1968

[2] F.R.Gantmacher,"Matrizentheorie", Springer Verl. Heidelberg, 1986

[3] P.Galko et alt.,"The Mean Power Spectral Density of Markov Chain Driven Signals", IEEE Trans.Inform.Theory, vol.IT-27, pp.746-754, Nov.1981

[4] G.Bilardi et alt.,"Spectral Analysis of Functions of Markov Chains with Applications", IEEE Trans.Commun., vol COM-31, pp.853-860, July 1983

[5] P.Lancaster, M.Tismenetsky,"The Theory of Matrices", 2nd Ed., Academic Press, Inc.,(London) Ltd. 1985

Quantization theory and EC-CELP advantages at low bit rates *

Majid Foodeei [1,2] and Eric Dubois [2,1]

Abstract – The goal of this work is to analyze the advantages of recently introduced entropy-constrained code-excited linear predictive (EC-CELP) quantization [1]. The analysis is at low rates and in comparison with other EC quantization schemes. Based on N-th order rate-distortion function (RDF), EC quantization theory, and empirical methods, *RDF memory gain* and *empirical space-filling gain* (dimensionality N) at low bit rates are defined and calculated. These gains categorize and help us analyze and compare the available coding gains for various EC coders for a given rate and delay (N).

EC-CELP and other EC quantizers EC-CELP addresses the problems associated with high-quality (near rate-distortion bound) quantization of sources with memory, operating at low bit rates, with minimal delay, and low complexity. The objectives are met through combining advantages of VQ, predictive coding (PC), and analysis-by-synthesis with merits of closed-loop entropy constrained (EC) codebook design (details in [1]).

Other EC schemes include EC scalar quantization (ECSQ) and vector quantization (ECVQ). For sources with memory, configurations which use a suitable memory removal technique, such as transform coding (TC) and PC, result in more efficient combination techniques with lower delay (dimension N). They include EC block transform quantization (EC-BTQ), EC-DPCM, and EC predictive VQ (EC-PVQ) [1]. All of the above EC coders (except EC-BTQ) can be shown to be special cases of EC-CELP (EC-CELP's better performance). We use a stationary first order Gauss-Markov (GM(1)) source for our comparisons and analysis. Fig. 1 shows the performance advantage of EC-CELP over other EC schemes for a given N. EC-BTQ results for a=0.9 are from a previously published work. For other a's, the values are predicted from ECVQ.

Coding gains analysis This analysis is based on RDF values ($\mathrm{SNR}_{\mathrm{RDF}}$), EC quantization theory, and empirical methods. Using a modified analysis scheme of [2], for low rates and general EC coders we define coding gains over basic ECSQ of *RDF memory* gain and *empirical space filling* gain. For a given dimension N and rate R, using a GM(1) source we have

$$\Delta^{\mathrm{GM(1)}}_{\mathrm{memory}}(N,R) = \mathrm{SNR}^{\mathrm{GM(1)}}_{\mathrm{RDF}}(N,R) - \mathrm{SNR}^{\text{i.i.d. Gaussian}}_{\mathrm{RDF}}(N,R)$$
$$\Delta^{\text{i.i.d. Gaussian}}_{\mathrm{filling}}(N,R) = \mathrm{SNR}^{\text{i.i.d. Gaussian}}_{\mathrm{ECVQ}}(N,R) - \mathrm{SNR}^{\text{i.i.d. Gaussian}}_{\mathrm{ECSQ}}(R).$$

For the low rate region the N-th order RDF is obtained parametricly. The top and bottom graphs in Fig. 2, show the memory and filling gains. For EC-CELP, the ideal PC effective N is high and hence should nearly provide the high N gains (top graph). The middle graphs show predicted memory gain for other coders. The analysis-by-synthesis feature of EC-CELP, in effect provides for intra-block PC gain (EC-CELP advantage over EC-PVQ). As the high-R ideal PC estimated PVQ gains in Fig. 2 show, loss of memory gain due to lack of intra-block PC gain could be substantial. An EC coder SNR is approximately the $\mathrm{SNR}^{\mathrm{Gaussian}}_{\mathrm{ECSQ}}$+ coder memory and filling gains. The combined memory and filling gains over ECSQ of EC-CELP for a given N and R is the highest. Hence it yields the highest SNR (Fig. 1). The efficient memory removal in EC-CELP allows for the concentration of VQ on the remaining memory redundancies (especially quantization) and the filling gain. Also since the resulting EC-CELP codebook size is not high the resulting EC-CELP complexity is also relatively low.

* This research was supported in part by a grant from the Canadian Institute for Telecommunications Research (CITR) under the NCE program of the Government of Canada.

[1] Electrical Engineering, McGill University, 3480 University Street, Montréal, PQ, CANADA H3A 2A7

[2] INRS-Télécommunications, Université du Québec, 16 Place du Commerce, Verdun, PQ, CANADA H3E 1H6

E-mail: {foodeei, eric}@INRS-Telecom.UQuebec.CA

References

[1] M. Foodeei and E. Dubois, "Entropy-constrained code-excited linear predictive quantization (EC-CELP)," in *IEEE Inter. Symp. Inf. Theory*, June 1994.

[2] T. D. Lookabaugh and R. M. Gray, "High-resolution quantization theory and the vector quantization advantage," *IEEE Trans. Inf. Theory*, vol. 35, pp. 1020–1033, Sept. 1989.

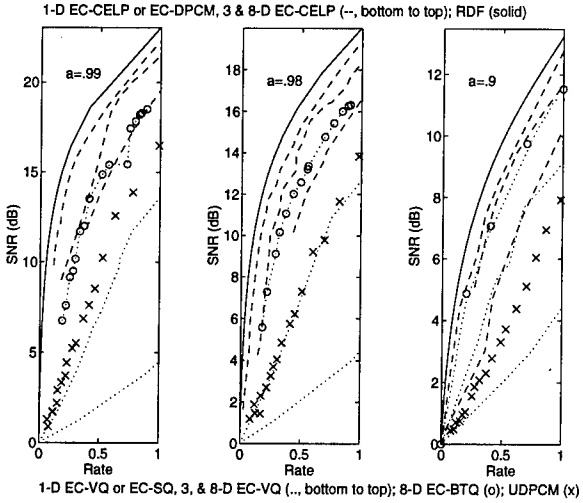


Fig. 1 Performance advantage of EC-CELP.

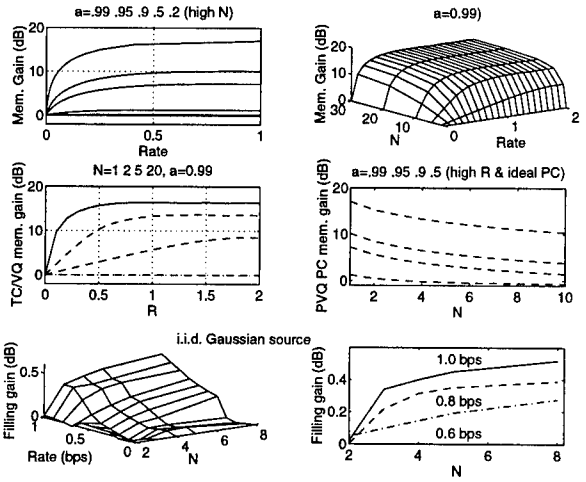


Fig. 2 Top graphs: High and low N RDF memory gains for Gauss-Markov source with coefficient a. Middle graphs: Analysis to show a comparison of theoretical memory gains between VQ/TC, PVQ, and CELP. Bottom graphs: low rate empirical space filling gains for i.i.d Gaussian source.

Neural Processing of Information

Robert L. Fry

The Johns Hopkins University/Applied Physics Laboratory, Laurel, MD 20723, USA

Abstract – **A model is proposed in which the neuron serves as an information channel. An application of the Shannon information measures of entropy and mutual information taken together in the context of the proposed model lead to the Hopfield neuron model with a conditionalized Hebbian learning rule and sigmoidal transfer characteristic.**

I. INTRODUCTION

A maximum entropy (ME) formulation is shown to provide the basic functional form of the model neuron including synaptic weights and a sigmoidal transfer characteristic. This formulation required the assumption of a set of measurement functions which are in turn a function of both synaptic inputs and neuron output. Furthermore, an ME formulation requires the specification of the statistical moments of the selected measurement functions which must somehow be supplied by an unspecified source. An ME formulation is underconstrainted in the sense that the model neuron cannot find a uniquely preferable set of moment constraints. Alternatively, a maximum mutual information (MMI) formulation is shown to be fully constrained in this regard and can make exclusive use of locally available information. Solutions take the form of the Hopfield neuron model with a requirement for a feedback learning methodology which takes the form of Hebbian learning whereby the synaptic weights are only modified in response to the conjunction of input and output neural events. A modification of an adaptation equation of Oja [1] provides for an algorithmic solution.

II. MAXIMUM ENTROPY (ME) FORMULATION

An ME formulation yields a Boltzmann distribution for a single neuron which extracts (measures) certain moments from its environment. A maximum likelihood decision rule results which corresponds to that of a deterministic Hopfield [2] neuron model. A stochastic decision rule is also possible which first requires the computation of an evidence function which can then be passed through a sigmoidal non-linearity. Described results require the specification of the moments of a set of N measurement functions by a nonspecific supervisor. This is considered undesirable regarding the development of useful computing structures constrained to use locally available information only.

III. MAXIMIZED MUTUAL INFORMATION (MMI)

Maximization of the mutual information between the neuron vector input $\mathbf{x}$ and output y can be accomplished if a ME distribution form is assumed for $P(\mathbf{x},y)$. The objective is to find the Lagrange set $\boldsymbol{\lambda}\epsilon\Lambda$ which maximizes the mutual information between $\mathbf{x}$ and y. This requires finding the extremum of an objective function $J(\boldsymbol{\lambda}) = I(\mathbf{x};y;\boldsymbol{\lambda}) + (\alpha/2)\boldsymbol{\lambda}^T\boldsymbol{\lambda}$ over some permissible $\Lambda \subset \mathbf{R}^N$ where the additional constraint $\boldsymbol{\lambda}^T\boldsymbol{\lambda}=\gamma^2$ is imposed. Without this constraint, an obvious extremum is $\boldsymbol{\lambda}=0$ in which case $I(\mathbf{x};y;\boldsymbol{\lambda})=0$. A derived Gibbs Mutual Information Theorem states that the extrema of $J(\boldsymbol{\lambda})$ can be found by solving a system of linear equations which lead to a conditionalized principal component analysis of the neural input. This results in a Hebbian learning rule analogous to biological models. This learning rule attempts to simultaneously minimize the conditional entropy of the output given the input $H(y \mid \mathbf{x})$ and also the entropy of the output $H(y)$ such that $P(y)=1/2$ implying that the neuron output has maximum entropy $H(y)$ for a one-bit channel. An extremely simple numerical algorithm serves to implement the developed strategy. Simulation results verify analytical derivations using simulated test data. These results indicate the model neuron automatically distinguishes input vectors into two equally probable classes based on degree of similarity. The biological equivalent of an action potential is generated for the preferred class.

REFERENCES

[1] Oja, E., "A simplified neuron model as a principal component analyzer," *J. of Math. Biol.* **15**, 267-273, 1982.

[2] Hopfield, J. J. and Tank, D. W., "Neural computations of decisions in optimization problems," *Biological Cybernetics*, 52, 141-152, 1985.

The Most Informative Stopping Times for Viterbi Algorithm: Sequential Properties

Joseph A. Kogan
Courant Institute of Mathematical Sciences, NYU, New York, NY 10012, USA

Abstract — **Sequential properties of the Viterbi algorithm are studied basing on a renewal sequence of the most informative stopping times which can be explicitly found during the Viterbi recognition of the most likeliest hidden Markovian state-sequence.**

I. Introduction

The Viterbi algorithm (VA) [1] allows to find the most likeliest state-sequence (MLSS) of a finite hidden Markov chain (HMC) $\{h_t\}$ indirectly observed through a process $\{z_t\}, t = 0, \dots, N$. An optimal rule for the VA can be found via maximizing the next additive criterion $\ln P\{\hat{h}_0^N, z_{-n}^N\}$ by a dynamic programming (DP) method [1]–[4]. Then the Viterbi recognition of $\hat{h}_0^{N-1}$ or the optimal segmentation of the observations z_0^{N-1} can be obtained by the backtracking $t = N-1, \dots, 0$: $\hat{h}_t \doteq k_{t+1}(h_{t+1})$, where $\hat{h}_N = \arg\max_{h_N} d(h_N)$ and $d(\cdot)$ is the corresponding additive functional for this DP problem.

The direct implementation of DP requires to store the values of $\hat{h}_t$ what fills up a table $\mathbf{K}(m \times N)$ with columns of back pointers $k_t : \hat{H}_t \longrightarrow \hat{H}_{t-1}, t = N, N-1, \dots$ with $\hat{H}_N = H = \{0, 1, \dots, m-1\}$ but if for a some moment $s, \exists j \in H$: $k_{s+1}(\hat{H}_{s+1}) \equiv j$ for all $\hat{h}_t \in \hat{H}_t = H, t > s$, then $\hat{h}_s \equiv j$ is called **a special column (SC)** in the table $\mathbf{K}$ of optimal DP decisions [2], [3].

Then the moments of the *SCs* appearing are **the most informative stopping times (MISTs)** for the Viterbi recognition of HMS [4] because after their appearing further observations don't change the previous decisions of the VA.

II. Main Results

1. The space of decision for the VA has the same structure as in Sequential analysis: the regions of acceptance hypotheses correspond to the regions of the SCs appearing as well as the region of continuation which is located in the middle. The bounds of these regions have a representation via $\min_i \ln p_{ji}/p_{li}$ [3], [4].

2. For a HMC with two states and the matrix of transition probabilities $P = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix}$, and τ_i there are three types of the back pointers decisions:

i) Identical decisions: $p + q > 1, A = \ln(1-p)/q, B = \ln p/(1-q)$. If $D_t^{10} = d_t(1) - d_t(0) \le A$, then $k_{t+1}(h_{t+1}) \equiv \hat{h}_{t=\tau_i} \equiv 0$, $k(h_{\tau_i - s}) = 0, s = 1, \dots, \tau_i - \tau_{i-1} - 1$. If $D_t^{10} \ge B$, then $k_{t+1}(h_{t+1}) \equiv \hat{h}_{t=\tau_i} \equiv 1$, $k(h_{\tau_i - s}) = 1, s = 1, \dots, \tau_i - \tau_{i-1} - 1$.

ii) Alternate decisions: $p + q < 1, A = \ln p/(1-q), B = \ln(1-p)/q$. If $D_t^{10} \le A$, then $k_{t+1}(h_{t+1}) = \hat{h}_{t=\tau_i} \equiv 0$, and

$$k(h_{\tau_i - s}) = \begin{cases} 1, & \text{if } s = 2r-1 \\ 0, & \text{if } s = 2r, \end{cases}$$

If $D_t^{10} \ge B$, then $k_{t+1}(h_{t+1}) = \hat{h}_{t=\tau} \equiv 1$, and

$$k(h_{\tau_i - s}) = \begin{cases} 0, & \text{if } s = 2r-1 \\ 1, & \text{if } s = 2r, \end{cases}$$

$r = 1, \dots, \tau_i - s < \tau_{i-1}$.

iii) Immediate decisions ($RCO = \emptyset$): $p+q = 1, A = B = 0$, (the underlying MC is degenerated into the independent trials and the *SCs* appear at each observation.)

Thus, (i) For $m = 2$ to store the intermediate information of back pointers is not necessary; (ii) One can get the same Viterbi recognition for different HMM; (iii) For $m \ge 3$ the VA can be analyzed as m-1 dimensional random walk on the underlying Markov chain with m states.

3. For an anticircle HMC with one ergodic class the SCs appears infinitely often a.s. and the mean and variance of the time of the SCs appearing can be estimated in many important cases via the analogues of the Wald's identities for random walk on a Markov chain.

4. As in the sequential analysis can estimate the error of Viterbi recognition [4]. If $P_i^\tau\{\tau(A,B) < \infty\} = 1\}, i = 0, 1$ and

$\alpha_\tau(A,B) < 1$, $\beta_\tau(A,B) < 1$, then $\ln^\alpha/(1-\beta) \le A, B \le \ln^{(1-\alpha)}/\beta$. But here, in duality to the sequential test of statistical hypotheses, the constants A and B are given.

5. The **duality** between the Wald's sequential analysis and the VA allows us also to represent the classical sequential problems such as testing of two simple hypotheses (TTSH) or change-point-distribution detection (CPDD) via the VA.

(i) $P = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}, 1 > \epsilon > 0$, for TTSH.

When ($\epsilon \longrightarrow 0$), the VA recognizes the true Markov state-sequence and therefore the true hypothesis with great accuracy. In this case the bounds of the region of observations tend to $\pm\infty$ as $\epsilon \longrightarrow 0$, so the first and second kinds of errors tend to 0.

(ii) $P = \begin{pmatrix} 1-p & p \\ \epsilon & 1-\epsilon \end{pmatrix}, 1 > p, \epsilon > 0$, for CPDD.

(iii) $P = \begin{pmatrix} \epsilon & 1-\epsilon \\ 1-\epsilon & \epsilon \end{pmatrix}$, for a Periodical chain.

6. The renewal properties of the MIST sequence can be used for the regenerative stochastic simulation for the VA and estimation of unknown parameters of a HMM by a segmental K-means recognition.

References

[1] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE* , vol. 61, pp. 268–278, 1973

[2] V. V. Motl', I. B. Muchnik, "Segmentation of structural curves by a dynamic programming method," *Automation and Remote Control* , No. 1, pp. 101–108, 1985

[3] J. A. Kogan, "Optimal segmentation of structural experimental curves by the dynamic programming method," *Automation and Remote Control* , No. 7, pt. 2, pp. 934–942, 1988

[4] J. A. Kogan, "Exact Viterbi recognition of hidden Markovian sequences via the most informative stopping times," submitted to *IMA Proceedings*

Modeling Gauss Markov Random Fields at Multiple Resolutions [1]

Santhana Krishnamachari and Rama Chellappa
Department of Electrical Engineering and
Institute for Advanced Computer Studies
Univ. of Maryland, College Park, MD 20742, USA

Abstract — **A multiresolution model for Gauss Markov random fields (GMRF) is presented. Based on information theoretic measures, techniques are presented to estimate the GMRF parameters of a process at coarser resolutions from the parameters at fine resolution.**

I. Introduction

There has been an increasing interest in using statistical techniques for modeling and processing images in the computer vision community. Most of the research has been restricted to Markov random field models, rightly so, because of the local statistical dependence of images. The main drawback of MRF techniques is that the associated optimization schemes are iterative and are usually computationally expensive. One way to reduce the computational burden is to use multiresolution techniques [1], [2]. In this paper, we present multiresolution models for Gauss Markov random fields.

II. Multiresolution Models

Let $\mathbf{\Omega^{(0)}} = \{(i,j) : 0 \leq i \leq M-1, 0 \leq j \leq M-1\}$ be a lattice on which a GMRF is defined. The superscript stands for the level in the image pyramid, $\mathbf{\Omega^{(0)}}$ is the lattice at the fine resolution and $\mathbf{\Omega^{(k)}}$ represents the lattice that is obtained by subsampling $\mathbf{\Omega^{(0)}}$, k times. The elements of $\mathbf{\Omega^{(\cdot)}}$ are indexed by s, where $s = (s_1, s_2)$. Let $X^{(k)}$ represent a random vector, obtained by ordering the random variables on the two-dimensional lattice $\mathbf{\Omega^{(k)}}$, through a row-wise scan . Let $X^{(0)}$ be modeled by a GMRF, then the joint probability density function of $X^{(0)}$ can be written as follows:

$$P^{(0)}(X^{(0)} = x) = \frac{exp\{-\frac{1}{2}x^T[\Sigma^{(0)}]^{-1}x^T\}}{(2\pi)^{\frac{M^2}{2}}(det\Sigma^{(0)})^{\frac{1}{2}}}$$

where $\Sigma^{(0)}$ is the covariance matrix of $X^{(0)}$. Equivalently, the process $X^{(0)}$ can be written in terms of a non-causal interpolative representation with a neighborhood $\eta^{(0)}$:

$$X_s^{(0)} = \sum_{r \in \eta^{(0)}} \theta_r^{(0)}(X_{s+r}^{(0)} + X_{s-r}^{(0)}) + e_s^{(0)}$$

where $e_s^{(0)}$, is zero mean, *spatially correlated* Gaussian noise with variance $[\sigma^{(0)}]^2$.

Hence a GMRF process can be completely characterized by the set of parameters $\{\theta, \sigma^2\}$. It can be shown that GMRFs lose Markovianity on subsampling resolution transformation. However, if lower resolution data are modeled by the exact non-Markov Gaussian measures, conventional optimization techniques based on Markov properties cannot be employed. We present two methods to estimate the parameters of Markov approximations at coarser resolutions. Let $P^{(k)}(X^{(k)})$ be the non-Markov pdf at kth resolution and $P_g^{(k)}(X^{(k)})$ be the family of Gauss Markov pdfs. Assuming a neighborhood $\eta^{(k)}$,

(1) a GMRF approximation can be obtained by minimizing $D[P^{(k)}(X^{(k)}) \;||\; P_g^{(k)}(X^{(k)})]$, where $D(.||.)$ is the Kullback-Leibler distance. It can be shown this computation is very similar to the conventional maximum likelihood estimation of the parameters except that instead of using sample covariances, this uses covariances calculated with respect to $P^{(k)}(X^{(k)})$ measure.

(2) a GMRF approximation can be obtained by minimizing $D[P^{(k)}(X_s^{(k)}/X_{s+r}^{(k)}) \;||\; P_g^{(k)}(X_s^{(k)}/X_{s+r}^{(k)})]$, $r \in \eta^{(k)}$. We call this *local conditional distribution invariance* approximation. It can be shown that this reduces to a form similar to the psuedo likelihood parameter estimation, again, uses covariances calculated with respect to $P^{(k)}(X^{(k)})$ measure. In this case, a closed form solution can be obtained for the parameters, but if the resulting parameters do not satisfy the positivity conditions [3], simple gradient descent method can be used.

Both methods presented above require covariance values $E_{P^{(k)}}(X_s^{(k)}X_{s+r}^{(k)})$, which can be computed given the GMRF parameters for $X^{(0)}$ as shown below:

$$\begin{aligned} X_s^{(k)} &= X_{2^k s}^{(0)} \\ E_{P^{(k)}}(X_s^{(k)}X_{s+r}^{(k)}) &= E_{P^{(0)}}(X_{2^k s}^{(0)}X_{2^k(s+r)}^{(0)}) \\ E_{P^{(0)}}(X_p^{(0)}X_q^{(0)}) &= \frac{1}{M^2}\sum_{s \in \Omega^{(0)}} \frac{(\lambda_{s_1}^{p_1}\lambda_{s_2}^{p_2})(\lambda_{s_1}^{q_1}\lambda_{s_2}^{q_2})}{1 - 2[\underline{\theta}^{(0)}]^T\phi_s} \end{aligned}$$

where $\lambda_i = exp(\sqrt{-1}\frac{2\pi i}{M})$.

We have used these models for multiresolution texture segmentation and have found that the multiresolution algorithm performs better than monoresolution algorithms with lesser computational requirement. In general, these multiresolution models can be applied for other low level image processing applications that use GMRF models.

[1] This work was supported in part by the National Science Foundation under Grant #ASC 9318183

References

[1] S. Lakshmanan and H. Derin, "Gaussian Markov Random Fields at Multiple Resolutions," in *Markov Random Fields: Theory and Applications* (R. Chellappa, ed.), pp. 131–157, Academic Press, 1993.

[2] B. Gidas, "A Renormalization Group Approach to Image Processing," *IEEE Trans. Patt. Anal. Mach. Intell.*, Vol. 11, No. 2, pp. 164–180, 1989.

[3] R. Chellappa, "Two-dimensional Discrete Gaussian Markov Random Field Models for Image Processing," in *Progress in Pattern Recognition* (L. N. Kanal and A. Rosenfeld, eds.), pp. 79–112, Elsavier, 1985.

Detecting Regularity in Point Processes Generated by Humans

Douglas Lake
Office of Naval Research, Arlington, VA 22217, USA

Abstract — **Detecting minefields in the presence of clutter is an important challenge for the Navy. Minefields have point patterns that tend to be regular for a variety of reasons including strategic doctrine, safety, tactical efficiency, and perhaps most intriguing the human element. For example, humans have a tendency to make lottery number selections, a one-dimensional discrete point process, in a non-uniform manner. In this paper, we introduce several simple procedures to detect regularity in point proceses.**

I. Introduction

The success of vital Navy amphibious assault operations depends on detecting minefields for subsequent neutralization or circumvention. Reconnaissance data from the surf zone can be modeled as a point process indicating locations where mines or more precisely minelike objects have been detected by a sensor. The presence of a minefield produces point patterns that tend to be regular (i.e., equally spaced). This property is a potentially valuable discriminant against natural clutter (such as rocks) that exhibit *complete spatial randomness* (CSR) indicative of a homogeneous Poisson process model (see [3]).

We first look at a simple and intuitive example. Lottery number selections consists of n different integers $x_1, x_2, \ldots, x_n$ between 1 and N inclusive. Proper characterization of human tendencies can dictate a strategy for selecting numbers that mitigate the probability of multiple winners and thereby effectively increases expected payoff. For example, it was shown in [2] that certain individual lottery numbers tend to be selected significantly more often than others. Presently, we focus on the interdependency between the entire sequence of n selected numbers.

II. Minimum and Maximum Gaps

Consider the distances or "gaps" between adjacent points $\{d_j = x_{j+1} - x_j : 1 \le j < n\}$. We hypothesize that humans tend to avoid extreme gaps because they seem "nonrandom". This translates into a disproportionately high frequency of selections without a low minimum gap and/or a high maximum gap. Moreover, we expect the *gap range*, the difference between the maximum and the minimum, to be small.

This approach was motivated by a simple but rather surprising recent observation [1] that randomly selected lottery numbers often have consecutive numbers. It is easy to show that for the minimum gap U

$$\Pr\{U > u\} = \binom{N - nu + u}{n} \Big/ \binom{N}{n} \qquad (1)$$

For example, the probability of no consecutive lottery numbers in Virginia where $N = 44$ and $n = 6$ is $\Pr\{U > 1\} = .462$. A straight-forward application of the inclusion-exclusion principle [5] applied to the maximum gap V gives

$$\Pr\{V \le v\} = \sum_{j \ge 0} (-1)^j \binom{n-1}{j} \binom{N - jv}{n} \Big/ \binom{N}{n} \qquad (2)$$

where the summation continues over positive entries. The null distribution for the gap range $W = V - U$ can also be found. For example, for the Virginia lottery $\Pr\{W \le 4\} = .03$ and $E[W] = 12$. Our conjecture suggests that the expected gap range is significantly smaller for human selections.

III. Minefield Detection Tests

Consider a point process of size n on a set A in R^2 that has been partitioned into N regions of equal area. Let M_r denote the number of regions containing exactly r points and Y_k denote the number of points in region k. If this process is generated by humans (i.e., minefields) the lottery analogy leads one to suspect less empty regions (smaller gaps) than under a CSR model.

The *empty boxes test* (EBT) based on M_0 has traditionally been used to detect the presence of too many empty boxes as an indication of lack of fit (see [4]). In these terms, humans tend to overfit. Dividing A into increasing number of regions and plotting the normalized EBT statistic at each scale produces a curve similar to the K-function (see [3]). However, the EBT approach can be more flexible and lacks edge effects and independence assumptions.

IV. Too Likely Likelihood Tests

The joint distribution of $Y_1, Y_2, \ldots, Y_N$ is multinomial under CSR. In particular,

$$\log f(y_1, y_2, \ldots, y_N) = \log n! - n \log N - \sum_{r=2}^{n} M_r \log r! \qquad (3)$$

so that even distributions of the points among the regions are more likely than uneven distributions. This seemingly creates a paradox with EBT. For example, the most likely value of (3) under CSR when $n = N$ corresponds to $M_0 = 0$ (one point in each region) for which EBT would reject CSR with the lowest possible p-value.

In this context, EBT is an example of a "too likely" likelihood test (TLLT). Without specifying an alternative, a TLLT rejects H_0 for high values of $\log f(T)$ where f is the null distribution for a statistic T. For the minefield detection scenario and large values of N, the mean and variances of the TLLT test statistic can be estimated using a Poisson approximation.

References

[1] Berman, D., *College Mathematics Journal*, vol. 25, pp. 45-47, 1994.

[2] Cover, T. and Stern, H., "Maximum Entropy and the Lottery,". *Journal of the American Statistical Association*, vol. 45, pp. 980-985, 1989.

[3] Cressie, N., *Statistics for Spatial Data*, New York: Wiley, 1991.

[4] Johnson N. and Kotz S., *Urn Models and their Applications*, New York: Wiley, 1977.

[5] Niven, I., *How to Count without Counting*, New York: Random House, 1965.

New Distortion Measures for Speech Processing

TA-HSIN LI* and JERRY D. GIBSON†
Texas A&M University, Collage Station, TX 77843

Abstract – **New distortion measures are derived from a recently proposed characterization function of stationary time series and are shown to be more robust than some commonly-used distortion measures such as the Kullback-Leibler spectral divergence in speech processing.**

I. INTRODUCTION

Distortion measures are widely used in speech processing to quantify the deviations of speech signals in correlation structure, and among the most successful ones is the Itakura-Saito (IS) distance of spectral densities [2], also known as the Kullback-Leibler information divergence [3]. Although in many cases the IS distance is quite effective in discriminating signals and detecting special changes, its lack of robustness is also well known documented in the literature, especially when the signals are mixtures of narrow and wide band components such as voiced speech waveforms (e.g., [1]). On the basis of a method called *parametric filtering*, we propose some new distortion measures that are shown to be more robust than the IS distance.

II. NEW DISTORTION MEASURES

Given a zero-mean stationary signal $\{X_t\}$, the parametric filtering method characterizes the correlation structure of $\{X_t\}$ by the demodulated first-order autocorrelation of the form

$$\gamma_\theta(\eta) := \Re\{e^{-i\theta}\rho(\alpha)\} \quad (-1 < \eta < 1),$$

where $\rho(\alpha)$ is the first-order autocorrelation of the filtered signal $X_t(\alpha) := \overline{\alpha}\, X_{t-1}(\alpha) + X_t$ with $\alpha := \eta e^{-i\theta}$. Among other interesting properties of $\gamma_\theta(\eta)$, it can be shown [4], [5] that $\gamma_\theta(\eta)$ uniquely determines the correlation structure of $\{X_t\}$ for almost any θ and is infinitely differentiable in $\eta \in (-1, 1)$ even for mixed-spectrum signals of which the spectral density does not exist. Using these properties, we define

$$\begin{aligned} p_\theta(\eta) \;:=\; & \tfrac{1}{2}\left[\gamma'_\theta(\eta) + (\gamma_\theta(\eta_a) + 1)\,\delta(\eta - \eta_a)\right. \\ & \left. + (1 - \gamma_\theta(\eta_b))\,\delta(\eta - \eta_b)\right], \end{aligned}$$

for any $-1 < \eta_a < \eta_b < 1$, where $\delta(\eta)$ is the Dirac delta. Clearly, the function $p_\theta(\eta)$ forms a (generalized) probability density in $[\eta_a, \eta_b]$ and, because of its equivalence to $\gamma_\theta(\eta)$, uniquely determines the correlation structure of $\{X_t\}$ for almost any θ. Therefore, we can define the following distortion measure using the Kullback-Leibler information divergence [3], namely

$$\begin{aligned} \kappa(p^0_\theta \| p^1_\theta) \;&:=\; \int_{\eta_a}^{\eta_b} p^0_\theta(\eta)\, K\left(p^1_\theta(\eta)/p^0_\theta(\eta)\right) d\eta, \\ \kappa(p^0_\theta; p^1_\theta) \;&:=\; \int_{\eta_a}^{\eta_b} K\left(p^0_\theta(\eta)/p^1_\theta(\eta)\right) d\eta, \end{aligned}$$

where $K(u) := u - \log u - 1$. Many other distortion measures can be defined, for instance, from the family of Renyi's information [6].

The IS spectral distance is known to be extremely sensitive to deviations of individual spectral peaks while less so to changes of overall spectral shapes (or envelopes). The new measures $\kappa(p^0_\theta \| p^1_\theta)$ and $\kappa(p^0_\theta; p^1_\theta)$ are potentially more robust than the IS distance because they are *finite* even when the spectral support changes. With this property, the new measures are able to avoid the disproportional sensitivity to frequency shifts and spectral peaks, and thus to discriminate correlation structures by treating the discrete and continuous components on an equal basis.

*T. H. Li is with the Department of Statistics.

†J. D. Gibson is with the Department of Electrical Engineering. He is supported by NSF grant NCR-93-03805.

REFERENCES

[1] R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. ASSP*, vol. 36, pp. 29–40, 1988.

[2] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and format frequencies," *Electron. Commun. Japan*, vol. 53-A, pp. 36–43, 1970.

[3] S. Kullback, *Information Theory and Statistics*, New York: Dover, 1968.

[4] T. H. Li, "Discrimination of time series by parametric filtering," Tech. Rep. 212, Dept. of Statistics, Texas A&M Univ., College Station, 1994.

[5] T. H. Li and J. D. Gibson, "Discriminant analysis of speech by parametric filtering," *Proc. 28th Conf. Inform. Sci. Syst.*, 1994.

[6] E. Parzen, "Time series, statistics, and information," in *New Directions in Time Series Analysis, Pt. I*, D. Brillinger *et al.* Eds., New York: Springer; pp. 265–286, 1992.

Nonparametric kernel estimation for error density

Zhu Yu Li and Shu Zhao Zou

Dept. of Math., Sichuan University, Chendu, China, 610064

Consider a linear model

$$y_i = x_i'\beta + e_i, \qquad i = 1,2,\cdots, \tag{1}$$

x_i's are $p(\geqslant 1)$dimension known vectors and $\beta(\in R^p)$is an unknown parametric verctor and e_i are assumed i. i. d. r. v.'s from a common unknown desity function $f(x)$ with

$$med\,(e_i) = 0 \tag{2}$$

Based on LAD(Least Absolute Deviations)estimator $\tilde{\beta}$ of β, we propose a nonparametric method to estimate unknown $f(x)$. A kernel estimator $\tilde{f}_n(x)$ is obtained as

$$\tilde{f}_n(x) = (nh_n)^{-1}\sum_{i=1}^{n} K(\frac{\tilde{e}_i - x}{h_n}), \qquad x \in R^1, \tag{3}$$

residuals $\tilde{e}_i = \tilde{y}_i - y_i$, h_n is a postive number, called as window width, $k(\cdot)$ is a Borel measurable function on R^1. Large sample properties of $\tilde{f}_n(x)$ are studied. Some computational esamples are also given.

References

[1]G. X. Chain, Z. Y. Li and H. Tian, "Consitent nonparametric estimation of error distributions iin linear mdoel", ACTA Math. Applicate Sinica, Vol. 7, pp. 245-256, 1991.

[2]G. X. Chai & Z. Y. Li, "Asymptotic theroy for estimation of error distribution in linear model", Science in China. Ser. A, English Ed. Vol. 36, pp 408-419, 1993.

[3]X. R. Chen, Z. D. Bai, L. C. Zhao and Y. H. WU, "Consistency of minimum L_1—norm estimates in linear model", Pitman Research Notes in Math., Ser. 258, pp. 249-260, 1992.

Neural Networks for Error Correction of Hamming Codes

O. Mayora-Ibarra, A. Gonzalez-Gutierrez and J.C. Ruiz-Suarez
Instituto Tecnologico y de Estudios Superiores de Monterrey, Campus Morelos.
Apdo Postal 99-C, Cuernavaca Morelos 62050, Mexico

Abstract — **A comparative analysis of three neural network models: Backpropagation (BPP), Bidirectional Associative Memory (BAM) and Holographic Associative Memory (HAM); and a classical method for error-correction is presented. Each method is briefly described, results are reported and finally some advantages are concluded.**

I. Introduction

Error correction is an important topic in any digital communication system. Classical methods for error-correction are usually based in Hamming distance techniques. The use of new technologies like neural networks is presented as an option for those who need to solve the error-correction problem in an alternative way.

II. Description

For the classical methods, the linear-block code is used. In this method, the capability for error-correction is a function of the Hamming distance in the sense that there exist a theoretical limit that could be reached for error-correcting depending only in the minimum Hamming distance between the codewords. The BPP method is designed as a multilayer feedforward net based in a supervised learning model. The net is trained with a predefined set of all the cases involved for the correction; in other words, for a (n,k) code, all its combinations of one error must be trained [1]. All these input-output pairs travel along the layers into the output layer and then are compared with the desired output value constructing an error signal for each output unit. In this moment, the error signals are backpropagated along the net. This process is repeated until a steady state is reached. Once trained the net, new patterns are introduced and a response is obtained.

The BAM consists in two layers of processing elements completely interconnected between them. In the BAM's architecture there are weights associated to the connections between processing elements forming a matrix. This matrix is used to obtain the recall of the information when new data are tested. BAM is capable of reconstructing noisy data. The bidirectional nature of the BAM occurs during the recall process[2]. Once trained the net, testing data are introduced to the BAM. This data are propagated along the two layers and an output is generated. The output is propagated backwards and the outcome is compared with the previus input. If no error exists between them, the recall obtained is the last output generated, otherwise the process is repeated until a steady state is reached. The convergence of the recall is waranteed with a Lyapunov function involved in the stability of the system. The theoretical limit for error-correction was reached with the BAM.

The HAM bases its operation in the principle of optical holography of "enfolding" information of different phase in a single plain [3]. The way this analogy occurs is clearly shown in the ability of the HAM to superimpose multiple stimulus-response associations onto the identically same correlation set representative of synaptic connections within the neuron cell in a complex number domain. The external field of accepted words in the alphabet is transformed to a complex plane by means of a sigmoidal function. This encoded data is set in a matrix representation for training the net. A new stimulus set is presented to the HAM for testing. The HAM calculates the minimum difference between the trained data and the tested one. A response is generated with the contribution of the difference between vectors and the closest desired output.

III. Results

Different results were obtained for the three neural networks used. The results obtained with the BPP net allowed us to decode the 90 per cent of the cases when testing the net with the predefined input patterns(128 patterns[1,4]).

For an specific (7,4) code with minimum Hamming distance equal to three, one error was corrected with a BAM trained with only 16 words allowed in the alphabet.

HAM's results show that, as in the BAM's case, the theoretical limit of one error corrected was reached with a minimum of 16 alphabet words trained for the (7,4) code. Its important to notice that the HAM also corrected more errors than the theoretical limit in 60 per cent of the cases.

IV. Conclusions

This work shows that neural methods employed for error-correcting presents an alternative for other error-correcting techniques with the advantage of its simplicity of programming and in its good correcting rates.

References

[1] Heindrich Norman. "Associative memory networks, fault-tolerance and coding theory." IEEE International Joint Conference on Neural Networks, 1992.

[2] Kosko B. "Adaptive bidirectional associative memories." Applied Optics, 1987, pp. 4947-4960.

[3] Souchek B. and Iris Group. "Fuzzy, Holographic and Parallel Inteligence: The Sixth Generation Breakthrough." John Wiley and Sons, 1992.

[4] Stefano A. Di, et al. "On the use of neural networks for Hamming coding." IEEE, Int. Sym. On Circuits and Systems, pp. 1601-1604, 1991.

Discussion of a Statistical Channel

Ira S. Moskowitz & Myong H. Kang
moskowit@itd.nrl.navy.mil, mkang@itd.nrl.navy.mil
Information Technology Division—CHACS: Code 5540, Naval Research Laboratory, Washington, DC 20375,USA

Abstract — **This paper deals with a new type of covert channel problem that arose when we designed a multilevel secure computer (MLS) system, using a quasi-secure, asynchronous, communication device called the *Pump*. We call this new type of covert channel a statistical channel. It is our hope to get feedback from experts who work in the intersection of information theory and statistics.**

I. Introduction

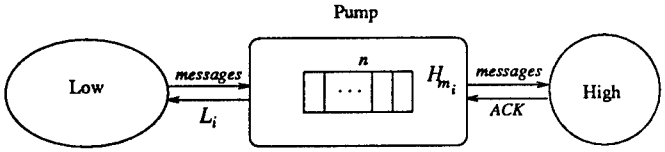
In a (MLS) system, Low may write to High, and High can read from Low, but High must never be able to write to Low. However, in a MLS system, the need for an acknowledgement (ACK), which is a write from High to Low, to a message sent by Low to High can violate the multilevel security policy by creating a covert (communication) channel.

Consider a case where Low sends messages to High. A simple approach that does not allow High to send an ACK to Low places a buffer between Low and High. Low submits messages to the buffer, the buffer sends the ACKs back to Low, and High then takes messages from the buffer. If the Low (sending) rate is faster than the High (receiving) rate, Low will write over unread data in the buffer (since the buffer is finite). An obvious solution to this problem is to not allow Low to send messages until there is a space in the buffer. This, however, results in a large capacity covert channel between High and Low (if Low is not allowed to send messages to a full buffer, then High can send symbols to Low by removing or not removing messages from the buffer and hence causing the buffer to be full or to have space on it).

II. The Pump

Our approach, the Pump [1], still places a buffer (size n) between Low and High, but has the buffer give ACKs at probabilistic times to Low based upon a moving average of the past m High response times (H_{m_i}). A high response time is the time from when the buffer tells High that it has a message to the time when High actually removes it. This has the double benefit of keeping the buffer from filling up and having a minimal negative impact upon performance.

Using a moving average is a very important part of the Pump. However, it gives rise to a new type of timing channel (for detail, see [1]). We will now sketch an implementation of the Pump. Let O_v be the communication overhead for the Pump. By this we mean that O_v is the minimum value for any L_i (which is the ith response to Low). The L_i are given by a random variable that has the density function $f_i(t)$.



There are two cases to discuss:

Case 1: The buffer is not full.

$$f_i(t) = \begin{cases} \alpha_i e^{-\alpha_i(t-O_v)}, & if\ O_v \le t, \\ 0, & otherwise. \end{cases}$$

The mean of the above density function is $O_v + 1/\alpha_i$. Since we wish for this mean of $f_i(t)$ to be equal to the moving average of the last m High ACK times (H_{m_i}) we see that $\alpha_i = 1/(H_{m_i} - O_v)$. If $H_{m_i} = O_v$, then set $1/\alpha_i = \epsilon$, a small number.

Case 2: The buffer is full.
This case is not germane to this paper.

III. Covert Channels

A timing (covert) channel exists when the output (Low) alphabet consists of the different times of the same response, these different times (e.g.,yes arriving at $3t$ or $5t$) being due to High behaviour. Historically, work on timing channels has used very simple tools from information theory, for example [2]. In the course of our work we have come upon a new type of timing channel that defies analysis by *our* research community. It is our hope that, by presenting a paper at this workshop, we will get feedback from experts who work in the intersection of information theory and statistics.

We introduce a new subspecies of timing channel referred to as a *statistical channel.* The Low alphabet consists of different time values and these time values are given by a random variable with certain parameters and these parameters are dependent upon High actions.

Definition 1 *If High can affect a parameter in the distribution of some system response time to Low, we say that there is a* **statistical channel** *between High and Low.*

In the Pump, High can modify the moving average by affecting the last m time values of High's responses to the Pump. It is possible for Low to detect differences in High's actions by trying to guess what the moving average is. This creates a statistical channel and, therefore, insecurity. For now, let us forget that the exponential density has been shifted by the communication overhead time, and simply view the inputs to the channel as the High response times. We state a simpler form of our problem as:

> *What is the capacity, in bits per unit time, of a communication channel where the output is an exponential random variable whose mean is the moving average of the past m input times?*

References

[1] Kang, M. H. and I. S. Moskowitz. "A pump for rapid, reliable, secure communication," Proceedings of the 1st ACM Conference on Computer and Communications Security, pp. 119–129, 1993.

[2] Moskowitz, I. S. and A. R. Miller. "The channel capacity of a certain noisy timing channel," IEEE Transactions on Information Theory, vol. 38, number 4, pp. 1339–1344, 1992.

Asymptotic Performance Evaluation of Mismatched Vector Quantizers Using Sub–Gaussian Sources

Frank Müller

Institut für Elektrische Nachrichtentechnik, RWTH, 52056 Aachen, Germany

Abstract — **Asymptotic (high rate) quantization theory is applied to the multivariate mismatch problem. This means, the question is addressed how much is lost if a vector quantizer which is matched to a specific source with given parameters is used for quantization of a source with different parameters. For parameterization of the sources sub-Gaussian processes are employed.**

I. Introduction

Vector quantization (VQ) is an often used method of lossy source coding. Usually, a vector quantizer is optimized by a training sequence which is expected to represent all the statistical characteristics of the samples to be quantized. In the most practical cases, however, it is impossible to find such a training sequence, because real source signals show time–varying statistics. Therefore, vector quantization is often implicitly coupled with the mismatch problem.

Performance evaluations of mismatched vector quantizers are interesting not only from an information theoretical point of view. They also give clues to design a robust quantizer under the knowledge that the actual source statistics are varying. If mismatch is eventually unavoidable, a vector quantizer should be designed under such conditions that mismatch around the operation point shall only weakly affect its optimum performance.

Only few results concerning mismatched vector quantizers are reported. The main reason being the lack of an appropriate comprehensive multivariate model. The situation changed when the engineering community became aware of the class of spherically invariant random processes (SIRP) and developed parametric source models [1][2]. SIRPs have the property that they are completely described by the univariate (marginal) density function and the linear statistical dependencies (covariances or covariations) between the random variables.

More important from a practical point of view, however, is that SIRP models reflect the statistics of a wide variety of sources. Band limited telephone speech samples [1], mean-removed image blocks [2], subband image statistics [3] and prediction error images [4] show elliptically shaped bivariate distributions, thus allowing SIRP modeling.

II. Mismatched Vector Quantization of Sub-Gaussian Sources

In [5] a new SIRP–model has been developed which employs *symmetric stable* densities [6] as marginal densities. Symmetric stable densities are defined as densities having a characteristic function of the form:

$$\phi(t) = \exp(-\gamma|t|^{\alpha}), \quad \text{with} \quad \gamma < 0,\ 0 < \alpha \leq 2. \qquad (1)$$

The stable distribution has much thicker tails than e.g. the Gaussian, thus allowing to model real world phenomena including outliers accurately with the aforementioned class of SIRPs. Moreover it has been shown in [5] that the class of SIRP–processes with symmetric stable densities is identical to a class of processes termed *sub–Gaussian processes* in mathematical statistics. Sub–Gaussian processes are completely parameterized by a shape parameter (called characteristic exponent) and a covariation matrix. With the characteristic exponent the shape of the distribution can be varied. The covariation matrix plays an analogue role as the covariance matrix in the classical second–order process theory. With the covariation matrix the variation (i.e. the stable analogue to the variance) as well as the linear dependencies between the samples can be adjusted.

Since the symmetric α-stable distribution (1) is completely specified by only two parameters (stable exponent α and variation γ) the mismatch problem can be formulated and solved in terms of *shape* and *variation* mismatch. Applying sub–Gaussian sources to the asymptotic (high rate — low distortion) quantization theory [7], we evaluate the relative performance of mismatched vector quantizers for these mismatch conditions. So, the question what happens, if the actual source distribution differs in shape from the distribution the quantizer is optimized for, can be answered employing sub-Gaussian processes as source model. It turns out that the robustness of a vector quantizer depends strongly on the dimension of the quantizer. Furthermore, vector quantizers respond — like scalar ones — unequally to mismatch around their operation point. However, the sensitivity against mismatch is reduced with increasing vector dimension.

References

[1] H. Brehm and W. Stammler, Description and Generation of Spherically Invariant Speech–Model Signals, *Signal Processing*, vol. 12, pp. 119-141, 1987

[2] Y. Du, "A Spherically Invariant Multivariate Distribution Model for Image Signals" (in German), *Archiv f. Elektronik und Übertragungstechnik*, vol. AEÜ-45, pp. 148–159, 1991

[3] F. Müller and C. Stiller. Multivariate modeling of subband image statistics using spherically symmetric distributions. *Proc. IEEE Int. Symp. on Inform. Theory*, p. 280, San Antonio, USA, Jan. 1993.

[4] F. Müller. *Multivariate statistical models for prediction error images in moving video.* Proc. International Workshop on Intelligent Signal Processing and Communication Systems (ISPACS 93), p. 297-302, Sendai, Japan, Oct. 1993.

[5] F. Müller and B. Hürtgen. *A new spherically invariant joint distribution model for image signals.* Proc. VIIth European Signal Processing Conference (EUSIPCO 94), in print. Edinburgh, Scotland, Sept. 1994.

[6] M. Shao and C. L. Nikias, Signal Processing with Fractional Lower Order Moments: Stable Processes and Their Applications, *Proceedings of the IEEE*, vol. 81, pp. 986-1010, 1993.

[7] A. Gersho, "Asymptotically Optimal Block Quantization", *IEEE Trans. on Inf. Th.*, vol. IT–25, pp. 373–380, 1979.

CONTINUOUSLY EVOLVING CLASSIFICATION OF SIGNALS CORRUPTED BY AN ABRUPT CHANGE

Thierry ROBERT & Jean Yves TOURNERET
ENSEEIHT/GAPSE, 2 rue Camichel, 31071 Toulouse Cedex, France
Phone : (33) 61588350 / fax : (33) 61588237 / email : robert@len7.enseeiht.fr

ABSTRACT - Bayes decision theory is based on the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known [1]. The aim of this paper is to show how blind sliding window AR modeling is corrupted by an abrupt model change and to derive a statistical study of these parameters.

I INTRODUCTION

The aim of this paper is to study the behaviour of continuously evolving classification when applied to signals presenting an abrupt change. AutoRegressive (AR) parameters are widely used to constitute the observation vector. The first part shows how sliding window AR modeling is corrupted when applied to signals that change abruptly. The second part studies the evolution of AR parameter statistics used in random process classification.

II SIGNALS PRESENTING ABRUPT CHANGE

Let us consider a particular signal $y(n)$ defined by the succession of two stationary signals $y_1(n)$ and $y_2(n)$ with an abrupt change occurring at $n = N_r$. A blind sliding window AR modeling of $y(n)$ (that is to say without any previously detected change) leads to the solving of a set of Yule and Walker equations [2]. The theoretical time dependent AR parameter vectors are given by :

$n \le N_r$ $\quad \underline{a}_n = \underline{\theta}_1$ (AR parameter vector of $y_1(n)$)

$N_r + 1 \le n \le N_r + L$ $\quad \underline{a}_n = [\underline{c}_n 0 \cdots 0]'$

$N_r + L + 1 \le n$ $\quad \underline{a}_n = \underline{\theta}_2$ (AR parameter vector of $y_2(n)$)

L being the AR model order and $\underline{c}_n$ a vector composed of $n - N_r - 1$ the Levinson Durbin Recursion (LDR) coefficients of the second signal $y_2(n)$. When $N_r \le n \le N_r + L - 1$, AR estimation then leads to an AR vector $\underline{a}_n$ with only $n - N_r - 1$ non zero coefficients [4].

III AR PARAMETER STATISTICS

We then study the case of random AR parameters. The $\underline{a}_n$ probability density function (p.d.f) allows us to analyse the evolution of the class shape when the sliding window moves. Let us denote :

$\underline{V}_k^T = [a_{(L,L)}, a_{(L-1,L-1)}, \ldots, a_{(k,k)}, a_{(k,k-1)}, \ldots, a_{(k,1)}]$

$a_{(i,j)}$ (with $j = 1, \ldots, i$) being the i^{th} order linear predictor coefficient estimator of the 2^{nd} model. k varying from 1 to L such that $n = N_r + 1 + k$.

For $n \ge N_r + L + 1$, we get $\underline{V}_k = \underline{\theta}_2$. The vector $\underline{V}_k$ is then the second AR model parameter vector, the p.d.f of which, denoted by $f_L(v_L, \ldots, v_1)$, is assumed to be known.

This last hypothesis is not restrictive because in pattern recognition, AR parameter statistics which characterises within-class scattering is usually assumed to be known (generally gaussian).

The next point of this study is to determine the $\underline{V}_{k-1}$ p.d.f, denoted by $f_{k-1}(v_L, \ldots, v_1)$ as a function of that of $\underline{V}_k$, k varying over $[1 \ldots L]$. The first $L - k$ components of these two vectors are equal. Their last k components verify the following relations :

$$1 \le j \le n - N_r - 1 \qquad a_{(n-1,j)} = \frac{a_{(n,j)} - a_{(n,n)} a_{(n,n-j)}}{1 - a_{(n,n)}^2}$$

These relations can be inverted and allow us to determine the jacobian of the transformation between the two vectors $\underline{V}_{k-1}$ and $\underline{V}_k$ [3]. We then obtain :

n odd :

$$f_{k-1}(v_L, \ldots, v_1) = (1 - v_k^2)^{\frac{k-1}{2}} f_k(v_L, \ldots, v_k, v_{k-1} + v_k v_1, \ldots, v_1 + v_k v_{k-1})$$

n even :

$$f_{k-1}(v_L, \ldots, v_1) = (1 + v_k)(1 - v_k^2)^{\frac{k-2}{2}} f_k(v_L, \ldots, v_k, v_{k-1} + v_k v_1, \ldots, (1 + v_k) v_{k/2}, \ldots, v_1 + v_k v_{k-1})$$

With L recursions, we may then determine the statistics of the different vectors $\underline{c}_n$ for $n = N_r + 1$ to $n = N_r + L$. These statistics allow us to study the evolution of the class shapes when the sliding window moves.

CONCLUSION

We show that sliding window AR modeling, applied to two stationary AR signals with an abrupt change, gives parameters which follow the Levinson Durbin Recursion. We give a recursive method making it possible to find the probability density function of these parameters when the sliding window moves. Class shapes may then be described in a continuously evolving classification.

REFERENCES

[1] K.FUKUNAGA "Statistical Pattern Recognition", academic press, 1990
[2] S.KAY "Modern Spectral Estimation : Theory and Application", Prentice Hall, 1988
[3] J.Y.TOURNERET " Etude Statistique des Coefficients de Reflexion ", GRETSI Juan-les-Pins, 1993.
[4] T.ROBERT & C.MAILHES "Autoregressive Estimation on Signals Presenting Abrupt Changes", EUSIPCO, Edindourgh, 1994

The Finite-Sample Risk of the k-Nearest-Neighbor Classifier under the L_p Metric

Robert R. Snapp[1]
Computer Science and Electrical Engineering Department
University of Vermont
Burlington, VT 05405 USA
snapp@emba.uvm.edu

Santosh S. Venkatesh
Department of Electrical Engineering
University of Pennsylvania
Philadelphia, PA 19104 USA
venkates@ee.upenn.edu

Abstract — **The finite-sample risk of the k-nearest neighbor classifier that uses an L_p distance function is examined. For a family of classification problems with smooth distributions in $\mathbb{R}^n$, the risk can be represented as an asymptotic expansion in inverse powers of the n-th root of the reference-sample size. The leading coefficients of this expansion suggest that the Euclidean or L_2 distance function minimizes the risk for sufficiently large reference samples.**

I. The k-Nearest-Neighbor Classifier

Let the elements of $\mathbb{L} = \{1, 2\}$ denote two states of nature, or pattern classes, and let P_1 and $P_2 = 1 - P_1$ denote their corresponding stationary prior probabilities. Each pattern is represented by a feature vector $\mathbf{X}$, drawn at random from $\mathbb{R}^n$. Specifically, patterns originating from class $\ell \in \mathbb{L}$ are generated by the stationary conditional distribution F_ℓ.

Labeled feature vectors are generated by a two-step process. First, a class $L \in \mathbb{L}$ is chosen at random so that $\mathbf{P}[L = \ell] = P_\ell$ for $\ell \in \mathbb{L}$; then a random feature vector is drawn according to F_L. After m independent repetitions of this process, we obtain the labeled reference sample,

$$\mathcal{X}_m = \{(\mathbf{X}^1, L^1), \ldots, (\mathbf{X}^m, L^m)\}.$$

Given an L_p metric, and an arbitrary point $\mathbf{x} \in \mathbb{R}^n$, the indices of the labeled feature vectors in $\mathcal{X}_m$ can be permuted so that

$$\|\mathbf{x} - \mathbf{X}^1\|_p \leq \|\mathbf{x} - \mathbf{X}^2\|_p \leq \cdots \leq \|\mathbf{x} - \mathbf{X}^m\|_p. \qquad (1)$$

Here $\|\mathbf{x}\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$ for $1 \leq p < \infty$, and $\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$, denote the L_p norm. The k nearest neighbors of $\mathbf{x}$ then form the subset $\{(\mathbf{X}^1, L^1), \ldots, (\mathbf{X}^k, L^k)\}$; and the k-nearest-neighbor classifier assigns $\mathbf{x}$ to class $L'(\mathbf{x}) = \mathrm{maj}(L^1, \ldots, L^k)$, viz., the most frequently appearing class label in the subset. (Ties, and degeneracies in (1), can be resolved by an arbitrary procedure.) Using this algorithm every point in $\mathbb{R}^n$ can be assigned to a class in $\mathbb{L}$.

II. The Finite-Sample Risk

Given a positive integer k, an L_p metric, and a finite random reference sample $\mathcal{X}_m$, a single test vector $(\mathbf{X}, L)$, drawn independently by the same random process, is assigned to class $L' = L'(\mathbf{X})$ by the k-nearest-neighbor classifier. We now consider the m-sample risk,

$$R_m = \mathbf{P}[L' = 1, L = 2] + \mathbf{P}[L' = 2, L = 1],$$

for two-class problems that satisfy the following smoothness conditions:

C1. For $\ell \in \{1, 2\}$, the class-conditional distributions F_ℓ are absolutely continuous over $\mathbb{R}^n$ and have corresponding densities f_ℓ.

C2. The mixture density, $f = P_1 f_1 + P_2 f_2$, is bounded away from zero a.e. over its probability-one support $\mathcal{S} \subset \mathbb{R}^n$.

C3. Each class-conditional density, f_ℓ, possesses uniformly bounded partial derivatives up to order $N + 1$ almost everywhere on its probability-one support.

C4. One or the other of the class-conditional densities vanishes close to the boundary of $\mathcal{S}$.

Theorem 1 *Under Conditions C1 through C4, there exist constants c_j, for $j = 2, 3, \ldots, N$, such that*

$$R_m = R_\infty + \sum_{j=2}^{N} c_j m^{-j/n} + O(m^{-(N+1)/n})$$

where R_∞ is the infinite-sample risk derived by Cover and Hart [1].

A proof of this theorem, including derivations of the leading coefficients, will be published separately. (An analogous proof for the nearest-neighbor classifier ($k = 1$) under the Euclidean metric ($p = 2$) appears in a recent paper [2].) For the coefficient c_2, we obtain

$$c_2 = D_n(p) \frac{\Gamma(k + 1 + \frac{2}{n})}{24 \left[\Gamma\left(\frac{k+1}{2}\right)\right]^2} \times \int_{\mathcal{S}} d\mathbf{x}\, f(\mathbf{x})^{1-\frac{2}{n}} (\widehat{P}_1 \widehat{P}_2)^{\frac{k+1}{2}} \left(\frac{1}{f_1} \nabla^2 f_1 + \frac{1}{f_2} \nabla^2 f_2 - \frac{2}{f} \nabla^2 f \right),$$

where

$$\widehat{P}_\ell = \widehat{P}_\ell(\mathbf{x}) = \frac{P_\ell f_\ell(\mathbf{x})}{f(\mathbf{x})}$$

denotes the posterior probability that a feature vector with value $\mathbf{x}$ originates from class ℓ. In the above,

$$D_n(p) = \frac{\Gamma\left(\frac{3}{p} + 1\right) \Gamma\left(\frac{n}{p} + 1\right)^{1+(2/n)}}{\Gamma\left(\frac{n+2}{p} + 1\right) \Gamma\left(\frac{1}{p} + 1\right)^3}$$

has a global minimum at $p = 2$ for fixed $n > 1$. This suggests that under the above assumptions the Euclidean metric is the optimal L_p distance function, if m is sufficiently large.

[1] This work was supported in part by Rome Laboratory, Air Force Material Command, USAF, under grant number F30602-94-1-0010.

References

[1] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21–27, 1967.

[2] D. Psaltis, R. R. Snapp, and S. S. Venkatesh, "On the finite sample performance of the nearest neighbor classifier," *IEEE Trans. Inform. Theory*, vol. IT-40, 1994, pp. 820–837.

Characteristics of a Statistical Fuzzy Grade-of-Membership Model in the Context of Unsupervised Data Clustering

Lisa M. Talbot*, H. Dennis Tolley†, Bryan G. Talbot‡, Harvey D. Mecham◇

* 11031 Barton Hill Ct., Reston, VA 22091

† Statistics Dept., Brigham Young Univ., Provo, UT

‡ The Analytic Sciences Corp., Reston, VA

◇ Chemistry Dept., Utah Valley State College, Orem, UT

Abstract — **We have elucidated the position of Woodbury's statistical fuzzy Grade-of-Membership (GoM) model in the unsupervised clustering domain. This implementation of the model is shown to operate not only on multivariate categorical data, but on permuted, or encoded, data as well.**

I. Introduction

We present results of a fuzzy unsupervised clustering paradigm, applying Woodbury's [1, 2] statistical fuzzy Grade-of-Membership (GoM) model to the problem of identifying natural clusters and statistical structure in data. Extensive theoretical development and empirical evaluation of the GoM clustering paradigm is presented by Talbot, et. al. [3].

II. GoM Clustering Paradigm

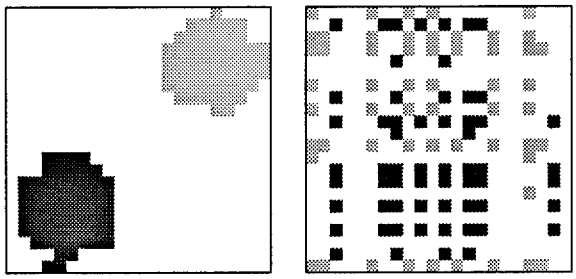
The GoM model simultaneously estimates profile probability densities and memberships for a fuzzy partition. Model parameters estimated from the data suggest a latent structure which may simplify coding, classification, and other analyses of high-dimensionality data. GoM clustering provides a more general framework for data analysis compared with conventional clustering paradigms in many cases. GoM model attributes that contribute to its generality in this context include its operation on categorical random variables, foundation in fuzzy set mathematics, and detachment from distance measures. A major model attribute, operation on categorical random variables, contributes to broad applicability—admitting categorical or even coded input data and allowing for nonlinear partitions. Because the data is represented by a finite alphabet, the model is also less sensitive to disparate scaling and outlying samples in many cases. A second model attribute, its fuzzy set basis, allows for characterization of more complex sources of heterogeneity in the data. A third attribute of the GoM model is its detachment from distance measures. By considering distance only indirectly through transitivity relationships, the model elucidates data structures primarily based upon characteristics of the estimated data distributions rather than upon distance computations between points in the space. This detachment from distance measures not only provides an unprecedented opportunity to evaluate the statistical composition of the data source but also offers new insights into structural mechanisms affecting coding performance.

III. GoM Clustering Examples

GoM clustering performance was compared with conventional vector quantization (VQ), fuzzy c-means (FCM) clustering, and deterministic annealing (DA) clustering to highlight differences between partitioning based upon distance measures versus that based upon statistical data structure. Continuous ordered data was quantized to produce categorical data.

Experimental outcomes for a two-dimensional unit step example demonstrate that the GoM model can provide an intuitively satisfying partition and structural determination as well as excellent background discrimination.

Figure 1 shows GoM clustering results for quantized and encoded multivariate Gaussian data derived from crisply defined distributions. The encoding clarifies the categorical nature of GoM clustering and also suggests potential applications for analysis of unconventional data sets which may be generated as the output of an encoder or classifier. In this case, GoM clustering provides an ideal partition of the encoded data as well as density estimates for sample data derived from each cluster. The log-likelihood value was experimentally shown to be a suitable clustering criteria, providing a strong correspondence to performance.



(a) (b)

Fig. 1: GoM clustering of multivariate Gaussian source: (a) quantized data and (b) quantized and encoded data. Gray level represents membership in each of two clusters.

IV. Conclusions

The GoM clustering paradigm supplements existing methods to broaden the application base and provide additional partitioning alternatives. The encouraging results suggest many applications in coding and classification, especially when employed in concert with conventional techniques.

References

[1] Max A. Woodbury and Jonathan Clive. Clinical pure types as a fuzzy partition. *Journal of Cybernetics*, 4(3):111–121, 1974.

[2] Kenneth G. Manton, Max A. Woodbury, and H. Dennis Tolley. *Statistical Applications Using Fuzzy Sets.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 1994.

[3] Lisa M. Talbot, H. Dennis Tolley, Bryan G. Talbot, and Harvey D. Mecham. Attributes of a fuzzy grade-of-membership model in an unsupervised clustering context. *IEEE Transactions on Fuzzy Systems*, 1994. Submitted for publication.

Function Estimation via Wavelets for Data with Long - Range Dependence

Yazhen Wang*
Department of Statistics, University of Missouri, Columbia, MO 65211, USA
Email: wang@stat.missouri.edu

Abstract – For a fractional Gaussian noise model, we derive asymptotics for minimax risks and show that wavelet estimates can achieve minimax over a wide range of spaces. This article also establishes a Wavelet - Vaguelette Decomposition (WVD) to decorrelate fractional Gaussian noise.

Introduction

Suppose we observe a function f from regression

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \cdots, n, \qquad (1)$$

where ε_i are zero mean stationary normal errors with long - range dependence.

Long - range dependence occurs in many applications. For example, it happens in data from geophysics and hydrology, economical time series, biological signals, image generation and interpolation, texture classification, noises in electronic devices, frequency variation in music, and burst error on communication channels. Signal processes with long - range dependence have much more persistent long term correlation structure than the well studied short - range processes such as ARMA processes and mixing processes. Traditionally, these process with long - range dependence have been mathematically awkward to manipulate. This has made the solution of many of the classical signal processing problems involving these processes rather difficult.

Fractional Gaussian noise provides a useful model for phenomenon exhibiting long - range dependence. We propose a fractional Gaussian noise model, which is an approximation of the nonparametric regression model (1), and then establish asymptotic results for minimax risks. Because of long - range dependence, the minimax risk and the minimax linear risk converge to zero at rates that differ from those for data with independence or short - range dependence. It is shown that a wavelet estimate with resolution level - dependent threshold can be "tuned" to achieve minimax over Besov bodies with $p \leq q$. Linear estimates can not achieve even the minimax rates over Besov classes when $p < 2$.

The key to prove the asymptotic results is to decorrelate fractional Gaussian noise and fractional Brownian motion via WVD by utilizing the idea of simultaneous diagonalization through WVD described in Donoho (1992) and the fact that Fractional Gaussian noise is linked to fractional differential operators which are almost diagonal in a wavelet basis.

Decorrelation of fractional Gaussian noise and fractional Brownian motion via WVD has its own interest. In fractal signal processing, it is very desirable to decorrelate fractional Gaussian noise and fractional Brownian motion (e.g. see Wornel and Oppenheim (1992)). Although wavelets reduce dependence of fractional Gaussian noise, the wavelet coefficients of fractional Gaussian noise and fractional Brownian motion are correlated and hence wavelets themselves do not decorrelate fractional Gaussian noise and fractional Brownian motion. Fractional Gaussian noise and fractional Brownian motion can be decorrelated by WVD.

Moreover, we employ two WVDs to solve the following linear inverse problems in the presence of indirect, noisy data with long - range dependence

$$y_i = (Kf)(x_i) + \varepsilon_i, \qquad i = 1, \cdots, n, \qquad (2)$$

where ε_i are zero mean stationary normal errors with long - range dependence and K is a linear transformation.

*This work was in part supported by NSF Grant DMS-94-04142

Unsupervised Medical Image Analysis by Multiscale FNM Modeling and MRF Relaxation Labeling

Yue Wang, Tülay Adalı, and Tianhu Lei
Department of Electrical Engineering, University of Maryland Baltimore County, Baltimore, MD 21228, USA

Abstract — **We derive two types of block-wise FNM model for pixel images by incorporating local context. The self-learning is then formulated as an information match problem and solved by first estimating model parameters to initialize ML solution and then conducting finer segmentation through MRF relaxation.**

I. Introduction

The main difficulty of unsupervised medical image analysis is that the model parameters are unknown and the priori context is unobservable. Any noncontextual algorithm is likely to perform poorly since locally there may not be sufficient information to make a good decision. The spatially dependence among pixels is one of some fundamental concerns and a reasonable assumption is that neighboring pixels are likely to have similar gray level and the same label. In the two main approaches to this problem, the MRF model-based techniques are often heuristically determined and computationally prohibitive [1], while the conventional FNM models only reflect partial context information in either global or pixel scale. This paper presents a new self-learning strategy based on stochastic regularization. The originalities are: 1) two types of block-wise FNM models are derived for pixel images by incorporating local context; 2) a unified information match criterion is applied to both model determination and pixel labeling.

II. Multiscale FNM Modeling

FNM modeling has proven to be a successful tool for medical image analysis that is mainly due to the validity of the independent approximation of pixels according to image statistics [2]. We extend this framework to include local context in multiscales. Assume a medical image with N^2 pixels and K regions. After dividing the image into disjoint blocks, the joint probability density function (pdf) can be well approximated by a block-wise conditional FNM model given by

$$P(\mathbf{x}) = \prod_{r=1}^{N^2/c^2} \prod_{k=1}^{K} [\prod_{i=1}^{c^2} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x_{ri}-\mu_k)^2}{2\sigma_k^2}\right)]^{l_{rk}} \quad (1)$$

where μ_k and σ_k^2 are the mean and variance of the kth region, c is the block size, and l_{rk} is the label associated with the rth block. By randomly reordering the neighboring pixels of the ith pixel, a new joint pdf of pixel images can be defined by introducing the local context into the standard FNM model in block form

$$Q(\mathbf{x}) = \prod_{i=1}^{N^2} \sum_{k=1}^{K} \left(\sum_{\hat{i}=1}^{c^2-1} \frac{l_{\hat{i}k}}{c^2-1}\right) \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right) \quad (2)$$

where $\hat{i}$ denotes the neighboring pixel with the label $l_{\hat{i}k}$, and the local context is naturally translated into non-parametric bindings by Bayesian priori probabilities. The problem addressed here is the combined estimation and detection of the regional (μ_k, σ_k^2) and structural (c, K) parameters and the contextual (l_{rk}, l_{ik}) variables, given the observations $\mathbf{x}$.

III. Information Criteria and Algorithms

The unsupervised estimation and detection can be characterized as an optimal information match problem. By minimizing a unified cost function, we solve this problem using stochastic regularization with two steps. Since parameters and variables in (1) are non-random unknown constants, we introduce a new model-fitting procedure derived from the modified global relative entropy, namely, the minimum bias/variance criterion (MBVC)

$$MBVC(K,c) = N^2 \sum_{u} Q_{\mathbf{x}}(u) \log \frac{Q_{\mathbf{x}}(u)}{Q(u|\hat{\mathbf{r}}_{ML}^{(K,c)})} + 3K - 1 \quad (3)$$

where $\hat{\mathbf{r}}_{ML}^{(K,c)}$ is the ML estimate of the parameter vector, $Q_{\mathbf{x}}(u)$ is the image histogram, and $Q(u|\hat{\mathbf{r}}_{ML}^{(K,c)})$ is the standard FNM. The balancing of decomposed model bias and variance yields

$$(K_0, c_0) = Arg\{\min_{K,c} MBVC(K,c)\} \quad (4)$$

with a simple optimal appeal: a minimum bias and variance model maximizes the information match [3]. Since (2) treats pixel-based labels as discrete random variables, by minimizing the expected Bayes risk, the Bayesian detection will classify pixel i into region j, if

$$j = Arg\{\max_{1\le k\le K} \left(\sum_{\hat{i}=1}^{c^2-1} \frac{l_{\hat{i}k}}{c^2-1}\right) \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right)\} \quad (5)$$

The block algorithms take advantages of the ML estimator being regional-structural separable and the MRF relaxation with local context revision consistency.

A. Multiple Resolution Block-Wise CM (MRBCM):

1. Given $\mathbf{r}^{(0)}$, $c = c_{max} + 1$
2. $c = c - 1$
   - $d_{KL}(r,j) = \log(\sigma_j/\sigma_r) + [\sigma_r^2 - \sigma_j^2 + (\mu_r - \mu_j)^2]/2\sigma_j^2$
   - $l_{rk}^{(m)} = 1$, if $k = Arg\{\min_{1\le j\le K} d_{KL}(r,j)\}$
   - $\mu_k^{(m+1)} = \sum_{r=1}^{N_b} l_{rk}^{(m)} \mu_r / \sum_{r=1}^{N_b} l_{rk}^{(m)}$, $\sigma_k^{2(m+1)} = \sum_{r=1}^{N_b} l_{rk}^{(m)} (\mu_r - \mu_k^{(m+1)})^2 / \sum_{r=1}^{N_b} l_{rk}^{(m)}$
   - Continue until $(\mathbf{l}^{(m+1)} - \mathbf{l}^{(m)}) = 0$
3. Stop when minimum global bias is reached.

B. Local Contextual Bayes Relaxation Labeling (LCBRL):

1. Given $\mathbf{l}^{(0)}$, m=0
2. m=m+1
   - Randomly visit each i and calculate $\sum_{\hat{i}=1}^{c^2-1} \frac{l_{\hat{i}k}}{c^2-1}$
   - Update l_{ik} according to (5)
3. Continue until $(\mathbf{l}^{(m+1)} - \mathbf{l}^{(m)}) = 0$.

Simulation results show the efficient and robust performance.

References

[1] R. C. Dubes and A. K. Jain, "Random field models in image analysis," *J. Appl. Statist.*, Vol. 16, No. 2, 1989.

[2] C. Bouman and B. Liu, "Multiple Resolution Segmentation of Texture Images," *IEEE T-PAMI*, Vol. 13, No. 2, 1991.

[3] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, 4, 1992.

An Additive Congruential Method for Generating a Multiple Occurrence Uniform Random Sequence

Dumitru M. Ionescu and Mark A. Wickert

Electrical and Computer Eng. Dept., Univ. of Colorado, 1420 Austin Bluffs Pkwy., Colorado Springs, CO 80933, USA

Abstract — **The formalism behind a novel additive congruential method is described. This method yields uniform random sequences whose outcomes occur more than once throughout the sequence.**

I. Introduction

An approach to generating a uniformly distributed pseudo-random sequence (PS) is presented; the method is based on addition rather than multiplication hence the name Additive Congruential Method (ACM). For a selected prime, p, the PS is a sequence of random variables (RV) over the Galois field $GF(p)$. The ACM yields a Markov PS, where each valid outcome appears more than once within the main period, and still obeys a uniform distribution (UD). It is argued that such a PS shows improved randomness over the standard Multiplicative Congruential Method (MCM) [1].

II. The Main Result

The PS is a chain of RVs over $GF(p)$, p prime. Theorem 1 and Corollary 1.1, stated, for the purpose of this paper, without proof, guarantee a uniform distribution on outcomes.

Theorem 1: Let p be a prime, $p > 2$, and let $k \in \aleph$, $1 \leq k \leq p-2$. Perform all possible modulo-p products between k distinct elements from $GF(p)\backslash\{0\}$. The number of occurrences, among the $\binom{p-1}{k}$ results will be the same for each element in $GF(p)\backslash\{0\}$ iff $k \left| \binom{p-1}{k-1} \right.$. □

The proof uses the structure of $GF(p)$ to write a finite difference equation with the distribution on outcomes as solution; its unique solution is the UD. Multiplication translates into addition if we let $x = \alpha^i$, $i = \overline{0, p-1}, \forall x \in GF(p)$, $\alpha \in GF(p)$ primitive. The corollary follows.

Corollary 1.1: Let p be a prime, $p > 2$, and let $k \in \aleph$, $1 \leq k \leq p-2$. Perform all possible modulo-p additions of k distinct integers from $GF(p)\backslash\{p-1\}$. The number of occurrences among the $\binom{p-1}{k}$ results will be the same for each element in $GF(p)\backslash\{p-1\}$ iff $k \left| \binom{p-1}{k-1} \right.$. □

The ACM relies on Corollary 1.1. If we calculate the empirical probability transition matrix (PTM), we see that it is doubly stochastic. Clearly, this is due to the UD on outcomes. Thus we may view the PS as a realization of some time invariant Markov process (TIMP) with a doubly stochastic PTM. A TIMP approaches a UD iff its PTM is doubly stochastic.

Definition 1: Model the ACM generated PS as a realization of some TIMP with doubly stochastic PTM. A natural measure of randomness for the ACM sequence is the degree of randomness of the associated Markov process.

In order to use Definition 1, we need to define a measure of randomness for the TIMP with doubly stochastic PTM. Such a measure, conjectured to be well-defined, is suggested by the following argument, which needs to be formalized. Contrary to the MCM where once an outcome has occurred an observer can count on the fact that it will NOT occur again within the main period, in the ACM PS there are multiple occurrences if $k > 1$. This is perceived as better randomness yet the distribution on the states of the TIMP is still not uniform, although the ensemble distribution is uniform. Consider the situations when an arbitrary integer in $GF(p)$ can be followed by (1) *some* integers in $GF(p)$, but *not* all of them and (2) *any* integer in $GF(p)$. Clearly, the latter PS is more random since we are less able to 'predict' the next outcome. It can be shown that if $\mathbf{P}$ is a doubly stochastic matrix of order m then $\mathbf{P}^n$, $n \in \aleph$ are stochastic and all entries in $\mathbf{P}_\infty = \lim_{n\to\infty} \mathbf{P}^n$ are equal. But $\mathbf{P}^n$ is the PTM of a Markov process described by $\mathbf{P}$ after decimating by n; hence retaining every nth sample achieves a more uniform distribution of states at any time instant. This is not possible with a MCM PS because of unique occurrences. The eigenvalues of $\mathbf{P}^n$ are the nth powers of the eigenvalues of $\mathbf{P}$ and approach the eigenvalues of $\mathbf{P}_\infty$, which are zero except for one that equals one. As stochastic matrices, $\mathbf{P}^n$, $n \geq 1$, and $\mathbf{P}_\infty$ each have a unity eigenvalue, thus the magnitudes of the non-unity eigenvalues of $\mathbf{P}^n$ are less than one (in order for the nth powers of the remaining $n-1$ eigenvalues to approach zero as in $\mathbf{P}_\infty$). Since p is prime, $m = p-1$ is even hence there is at least one more real eigenvalue $\lambda_{0n} \neq 0$, $|\lambda_{0n}| < 1$, for each of $\mathbf{P}^n$, $n \geq 1$ and the tendency to improve randomness as $n \to \infty$ reflects the tendency of λ_{0n} to approach zero. A randomness measure for a TIMP with doubly stochastic PTM $\mathbf{P}$ could be the inverse of the largest of the magnitudes of all real, non-unity, eigenvalues of $\mathbf{P}$. This however is impractical since $\lim_{n\to\infty} \lambda_{0n}^{-1} = \infty$. If for $\mathbf{P} = [p_{ij}]$, $\mathbf{P}_{adj} = [a_{ij}]$, $i, j = \overline{1, n}$ where $a_{ij} = p_{ij}$ if $p_{ij} = 0$ and $a_{ij} = 1$ if $p_{ij} \neq 0$, then a more attractive measure is $\Re = \max_i \{\lambda_i \,|\, |\lambda_i| < 1 \,, \exists \mathbf{x}_i \ni \mathbf{P}_{adj}\mathbf{x}_i = \lambda_i \mathbf{x}_i\} \leq p-1$. *This is well defined if as conjectured, $\Re$ increases as $\mathbf{P}_{adj}$ becomes less sparse.*

III. Example

For $p = 11$, $k = 3$ one implementation of the ACM yields PS: 03670369267036947123921458581470503690925818149258369 47036920379258147069258258581470381473614702570364692 58136925814147. The (doubly stochastic) PTM is

$$\mathbf{P} = \frac{1}{12}\begin{bmatrix} 0 & 0 & 1 & 8 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 9 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 9 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 8 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 10 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 2 & 0 & 8 \\ 9 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 8 & 1 & 1 & 0 & 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 9 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and $\Re = 4.07$. Interleaving every 30th entry until all entries are exhausted yields a PS with $\mathbf{P}_{adj}$ less sparse and $\tilde{\Re} = 6.48$.

References

[1] D. E. Knuth, *The Art of Computer Programming*, vol. 1, Addison-Wesley Publishing Co., 1968.

An Asymptotic Property of Model Selection Criteria

Yuhong Yang

Department of Statistics, Yale University, PO Box 208290, Yale Station, New Haven, CT 06520, USA

Abstract — **Probability models are estimated by use of penalized likelihood criteria related to AIC and MDL. The asymptotic risk of the density estimator is determined, under conditions on the penalty term, and is shown to be minimax optimal. As an application, we show that the optimal rate of convergence is achieved for density in certain smooth nonparametric families without knowing the smooth parameters in advance.**

I. Introduction

Both AIC [1] and MDL [2] are widely used model selection criteria based on information-theoretic considerations. Recent work described in the talk by Barron at this workshop suggests that in certain cases the minimum description length principle can yield a minimax optimal criterion of the form -log(likelihood)+const · m as opposed to -log(likelihood) + $\frac{m}{2}\log(n)$ where m is the number of parameters in the model and n is the sample size. The penalty term in this criterion is of the same order as that in AIC which takes the form -log(likelihood)+m. Previously an asymptotically optimal property was obtained for AIC applied to sequence of linear models in estimating a nonparametric regression function with fixed design [3][4]. In this work, we consider criteria of the form

$$-\sum_{i=1}^{n}\log f_k(X_i,\hat{\theta}_k)+\lambda_k m_k$$

where λ_k is a positive constant and $\hat{\theta}_k$ is the maximum likelihood estimator of θ_k in model k. In this work λ_k is specified so that the desired asymptotic results hold. Here $X_1,...,X_n$ are an i.i.d. sample from an unknown density $f(x)$ w.r.t. some σ-finite measure.

To handle also selection problem involving large numbers of models of each dimension m_k, we consider criteria of the form

$$-\sum_{i=1}^{n}\log f_k(X_i,\hat{\theta}_k)+\lambda_k m_k+C_k \qquad (*)$$

where C_k is a model complexity satisfying Kraft's inequality $\sum_k 2^{-C_k}\le 1$. We note however that $\lambda_k m_k$ does not necessarily correspond to a description of estimated parameters, so (*) does not necessarily have a total description length interpretation, so that the work of Barron and Cover [5] does not apply.

We evaluate the new criteria by comparing the Hellinger loss $d_H^2(f,f_{\hat{k},\hat{\theta}_{\hat{k}}})$ with an index of resolvability. The concept of resolvability was introduced in [5]. It naturally captures the capability of estimating an unknown function by a sequence of models. The index of resolvability can be defined as

$$R_n(f)=\inf_k\{\inf_{\theta_k\in\Theta_k} d_H^2(f,f_{k,\theta_k})+\frac{m_k}{n}+\frac{C_k}{n}\}\quad .$$

The first term $\inf_{\theta_k\in\Theta_k} d_H^2(f,f_{k,\theta_k})$ reflects the approximation capability of the model k to the true function $f(x)$, the second term $\frac{m_k}{n}$ reflects the variation due to estimating the best parameters in the model, and the last term $\frac{C_k}{n}$ reflects the complexity of the model relative to the sample size. The index of resolvability quantifies the best tradeoff among the approximation error, the estimation error and the model complexity.

II. Main results

It is shown in this work that with the new criteria and under some reasonable smoothness conditions on the parametric families and under some restriction on λ_k, the Hellinger loss $d_H^2(f,f_{\hat{k},\hat{\theta}_{\hat{k}}})$ is bounded in probability by the index of resolvability $R_n(f)$. With some additional conditions, the risk $Ed_H^2(f,f_{\hat{k},\hat{\theta}_{\hat{k}}})$ is proved to be bounded by a multiple of the index of resolvability $R_n(f)$, i.e.,

$$Ed_H^2(f,f_{\hat{k},\hat{\theta}_{\hat{k}}})=O(R_n(f))$$

As a consequence, by examining the index of resolvability for various nonparametric class of functions, the convergence rates of the modified AIC procedure can be easily upper-bounded. For some cases, the optimal rate of convergence is shown to be achieved.

III. An application

As an application, we consider estimating a density function on [0,1] using a sequence of exponential families with spline basis functions. The logarithm of the density is assumed to be in the Sobolev space W_2^s (which consists of all the functions on [0,1] having s square-integrable derivatives) with s unknown. The new criterion is used to select the spline order and the number of knots. For each s and each number of knots, separate spline models are considered for each radius constraint $\|\log f(x,\theta)\|_\infty\le r$, r=1,2,... .The correspoinding λ_k depends on r and s. We conclude that the optimal rate of convergence is achieved simultaneously for density function f with $f\in W_2^s$ for all s without knowing it in advance.

Acknowledgements

The author is grateful to his advisor Andrew Barron, whose guidance made the paper possible.

References

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," *in Proc. 2nd Int. Symp. Info. Theory*, 1972

[2] J. Rissanen "Universal coding, information, prediction, and estimation ," *IEEE Trans. on Information Theory*, vol. 30, 1983

[3] K.C. Li, "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set," *Ann. Statistics*, vol. 15, no. 3, 1987

[4] Y. Yang, "Complexity-based model selection, " prospectus submitted to Department of Statistics, Yale University, 1993

[5] A.R. Barron and T.M. Cover, "Minimum complexity density estimation, " *IEEE, Trans. on Information Theory*, vol. 37, no. 4, 1991

WAVELET NETWORKS FOR FUNCTIONAL LEARNING

Jun Zhang

Dept. EECS
University of Wisconsin-Milwaukee
Milwaukee, WI 53201
junzhang@ee.uwm.edu

Gilbert G. Walter

Dept. of Mathematics
University of Wisconsin - Milwaukee
Milwaukee, WI 53201
ggw@convex.csd.uwm.edu

ABSTRACT

A wavelet-based neural network is described. The network is similar to the radial basis function (RBF) network, except that the RBF's are replaced by orthonormal scaling functions. It has been shown that the wavelet network has universal and L^2 approximation properties and is a consistent function estimator. Convergence rates, which avoid the "curse of dimensionality," are obtained for certain function classes. The network also compared favorably to the MLP and RBF networks in the experiments.

1. INTRODUCTION

Recently, neural networks have become a popular tool in non-parametric function learning. While the multi-layer perceptron (MLP) is probably the most frequently used, its training process often converge too slowly or settle in undesirable local minima. The radial basis function (RBF) network can be trained more easily provided that certain network parameters (e.g., the centers and the variances) are properly preset. As a function representation scheme, the RBF network uses a family of locally-supported basis functions, which allows it to represent a "rich" class of functions. However, since the basis functions are generally non-orthogonal, the RBF representation is not unique (coefficients "harder to learn") and not the most efficient.

In this work, we replace the basis functions in the RBF network by an orthonormal basis, namely, the scaling functions associated with a orthonormal wavelet basis. For a given function, this "wavelet network", provides a unique (coefficients "easy to learn") and efficient representation. The use of orthonormal scaling functions also facilitates the theoretical analysis the network, such as universal approximation and consistency. The idea of using orthonormal wavelets in neural networks has also been investigated recently Zhang and Benveniste and by Pati and Krishnaprasad (see the reference list of [1]), who use non-orthogonal wavelets, and by Boubez (see also [1]), who is concerned more with classification than with function learning.

2. WAVELET NETWORKS

In this section, we briefly summarize the main theoretical and experimental results related to the wavelet network. More details can be found in [1]. For the sake of simplicity, we first look at the 1-D case (one input and one output).

Since in most practical applications, the function of interest has finite support, we assume that, without lose of generality, that $f(t) \in L^2(\mathbf{R})$ and has finite support. Let $g(t)$ be a wavelet-based approximation to $f(t)$. Then, there exists a sufficiently large M, such that

$$f(t) \simeq g(t) = \sum_k c_k \varphi(2^M t - k)(t). \qquad (1)$$

where $\varphi(t)$ is a compactly-supported (or fast-decaying) scaling function and k runs through a finite set of integers. $g(t)$ can be implemented as a three-layer network [1] and c_k's can be estimated by minimizing the mean square error between $f(t)$ and $g(t)$ over a training data set. Since multi-dimensional scaling functions can be obtained easily from $\varphi(t)$, the extension of the network to dimensions higher than one is straightforward.

The theoretical results related to the wavelet network are described by the following three theorems:

Theorem 1. The wavelet network has the properties of universal approximation and L^2 approximation.

Theorem 2. The rates of convergence for Theorem 1 can be made arbitrarily rapid in the following sense: for any $\alpha > 0$, there is a Sobolev space $\mathbf{H}_\beta$ such that for any $f \in \mathbf{H}_\beta$, there exists a sequence of wavelet networks f_n, where $n = 2^M$, such that

$$||f - f_n||_u = O(n^{-\alpha}), \qquad ||f - f_n||_{L^2} = O(n^{-\alpha}). \qquad (2)$$

Here $||\cdot||_u$ and $||\cdot||_{L^2}$ are the *sup* and L^2 norms, respectively.

Theorem 3. Assume that the training data are i.i.d. and uniformly distributed. Then, the wavelet network is L^2 consistent in the mean square sense and the rate of convergence for the coefficients is $O(1/N)$, where N is the size of training data set.

The proof of these theorems can be found in [1]. In the experiments, the wavelet network performed better than the MLP with similar complexity in learning discontinuous functions and the performance of the RBF became comparable to the wavelet network only when some of its parameters are preset according to the wavelet network.

3. REFERENCES

[1] J. Zhang, G. G. Walter, Y. Miao, and W. N. Lee, "Wavelet neural networks for function learning," submitted to IEEE Trans. Signal Processing.

(Late submission - Poster Session I)

A Comparison of Algorithms for Lossless Data Compression Using the Lempel-Ziv-Welch Type Methods

Adrian Traian Murgan, Radu Radescu
Applied Electronics Dept., Univ. "Politehnica" of Bucharest, 1-3 Armata Poporului Bld., RO

Abstract - **Lempel-Ziv-Welch methods and their variations are all based on the principle of using a prescribed parsing rule to find duplicate occurrences of data and encoding the repeated strings with some sort of special code word identifying the data to be replaced. This paper includes a general presentation of five existing lossless compression methods used in any application of digital signal processing. The comparisons are made experimentally by computer simulation.**

I. INTRODUCTION

The purpose of this paper is to compare the compression performances of five lossless compression algorithms applied to various types of files.

II. PRESENTATION OF ALGORITHMS

Algorithm 1 [1], [3] is the Lempel-Ziv 1977 method, Algorithm 2 [2] is the Lempel-Ziv 1978 method, Algorithm 3 [1], [5] is an intermediate 1992 method of Algorithms 1 and 2, Algorithm 4 [5] is the Welch variant of Algorithm 3 and Algorithm 5 [5] is an intermediate method of LZW method [4] and Algorithm 4. Note that the LZW method is a practical Welch type variant of Algorithm 2.

III. SIMULATION RESULTS AND REMARKS

In order to investigate the performance of the practical schemes, the proposed algorithms have been implemented by experimental computer programs, which were tested against various kinds of byte-oriented data. In addition to the compression ratio R(n), the size S(n) of the corresponding string table is shown in the table as a function of the input length n.

n [bytes]		100	200	500	1000	2000	5000	10000
Alg.	R(n)	.72	.65	.56	.50	.41	.35	.32
4	S(n)	355	451	748	1250	2247	5252	10254
Alg.	R(n)	.75	.67	.58	.52	.46	.39	.35
5	S(n)	321	384	625	809	1627	3074	5103
LZW	R(n)	.80	.72	.64	.56	.51	.42	.38
meth.	S(n)	298	370	420	701	996	1675	2982

In order to study the asymptotic convergence of the compression ratio, we used: a Turbo Pascal file of length 6250 bytes, a maximum length of 32 bytes for the source words, a length of 3 bytes for the code words and various values for the encoder buffer length. The results are:

nb [bytes]		100	200	400	800	1600
Algorithm 1	R(n)	76	.57	.42	.37	.33
Algorithm 4	R(n) =	.42				
Algorithm 5	R(n) =	.45				
LZW meth.	R(n) =	.51				

The algorithms have also been tested on different types of program and text files. The values obtained for the compression ratio R(n) are shown in the following:

File	Type	Size [bytes]	Alg. 4	Alg. 5	LZW
# 1	Pascal	2993	.50	.55	.60
# 2	Text	1237	.77	.81	.87
# 3	Pascal	6250	.42	.45	51
# 4	Pascal	10423	.32	.35	.38

The best compression ratio is given by Algorithm 4. In general, Algorithm 4 shows between 10 an 16 percent improvement over the LZW method, and Algorithm 5 shows between 7 and 11 percent improvement over LZW. Good values for the compression ratio are obtained only for input sequences with great length. For short-length files with small entropy, Algorithm 1 is the best. Also, regarding the memory space for encoding, Algorithm 1 has the best performances, because the other ones requires greater memory space for developing the string tables.

REFERENCES

[1] J. Ziv, A. Lempel, "A Universal Algorithm for Sequential Data Compression", IEEE Trans. Inform. Theory, vol. IT-23, no. 3, pp. 337-343, May 1977.

[2] J. Ziv, A. Lempel, "Compression of Individual Sequences via Variable-Rate Coding", IEEE Trans. Inform. Theory, vol. IT-24, no. 5, pp. 530-536, September 1978.

[3] G. G. Langdon, Jr., "A Note on the Ziv-Lempel Model for Compressing Individual Sequences", IEEE Trans. Inform. Theory, vol. IT- 29, no. 2, pp. 284-287, March 1983.

[4] T. A. Welch, "A Technique for High Performance Data Compression", IEEE Computer, vol. 17, pp. 8-19, June 1984.

[5] H. Yokoo, "Improved Variations Relating the Ziv-Lempel and Welch-Type Algorithms for Sequential Data Compression", IEEE Trans. Inform. Theory, vol. 38, no. 1, pp. 73-81, January 1992.

AUTHOR INDEX